UNIVERSIDADE DE SANTA CRUZ DO SUL

CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE SOFTWARE, ÊNFASE EM SOLUÇÕES DE GOVERNO.

Rodrigo de Abreu Batista

CLASSIFICAÇÃO AUTOMÁTICA DE CÓDIGOS NCM UTILIZANDO O ALGORITMO NAÏVE BAYES

Porto Alegre, janeiro de 2017.

Rodrigo de Abreu Batista

CLASSIFICAÇÃO AUTOMÁTICA DE CÓDIGOS NCM UTILIZANDO O ALGORITMO NAÏVE BAYES

Trabalho de conclusão apresentado ao Curso de Pós-Graduação em Engenharia de Software, ênfase em Soluções de Governo da Universidade de Santa Cruz do Sul - UNISC.

Orientadora: Ms. Daniela Bagatini

Porto Alegre, janeiro de 2017.

Rodrigo de Abreu Batista

CLASSIFICAÇÃO AUTOMÁTICA DE CÓDIGOS NCM UTILIZANDO O ALGORITMO NAÏVE BAYES

Este trabalho de conclusão foi submetido ao Curso de Pós-Graduação em Engenharia de Software, ênfase em Soluções de Governo da Universidade de Santa Cruz do Sul - UNISC.

Ms. Daniela Bagatini
Professora Orientadora

AGRADECIMENTOS

Agradeço à minha família pelo apoio incondicional e por sempre acreditarem em mim. Agradeço à minha mãe, Maria Luzia, ao meu pai, Mario Augusto e à minha namorada, Aline von Mühlen, por compreenderem a minha ausência e estarem incentivando que eu dê mais esse passo em direção ao meu crescimento profissional. Agradeço aos meus avós Alzírio, Loeci, Therezinha e Mário Jurandir, que hoje está no oriente eterno, pela educação e cuidado que sempre me proporcionaram, fazendo com que eu sempre estivesse no caminho do bem. Agradeço à professora Daniela Bagatini, por ter me orientado nesse trabalho, ao Programa de Pós-Graduação em Computação da UNISC, por fornecer uma educação de qualidade, à Procergs, por ter me possibilitado cursar essa Pós-Graduação, e a todos os meus colegas do curso de Pós-Graduação em Engenharia de Software com Ênfase em Soluções de Governo Eletrônico, pelas sugestões apontadas, que contribuíram para o desenvolvimento das ideias aqui apresentadas. Todos foram fundamentais para que esse trabalho fosse concluído.

“As you think, so shall you become.”

—Bruce Lee

RESUMO

Esse trabalho consiste em um estudo sobre a construção de um classificador de texto para a categorização automática de descrições de itens de produto em seus códigos NCM adequados. A construção desse classificador foi realizada utilizando a abordagem conhecida por aprendizado supervisionado. O treinamento desse classificador consistiu em três conjuntos de dados, o primeiro desses formado por itens de transações de NCM do Capítulo 20 (Bebidas, líquidos alcoólicos e vinagres), emitidas por uma mesma rede de varejo; o segundo composto por itens de transações também do Capítulo 20, porém pertencendo a estabelecimentos distintos; e o terceiro conjunto, formado por itens de transações pertencentes ao Capítulo 90 (Instrumentos e aparelhos de óptica, de fotografia, de cinematografia, instrumentos e aparelhos médico-cirúrgico, etc.) e emitidas por estabelecimentos distintos. Cada um desses conjuntos de dados possuem dificuldades distintas, foram previamente extraídos da base de Notas Fiscais do Consumidor (NFC) da Secretaria da Fazenda do Estado do Rio Grande do Sul e devidamente rotulados. O treinamento do classificador exigiu que os documentos fiscais, entradas para o classificador, fossem tratados utilizando técnicas de processamento de texto, com o objetivo de transformá-los no formato adequado para a calibragem do modelo. Durante a fase de pré-processamento foi realizada a tokenização, remoção de *stop words*, *stemming* e extração de vetor de características utilizando a medida TF-IDF. Ao final do processo, avaliou-se o desempenho do modelo baseada na metodologia de validação cruzada com 10-*folds*. Os resultados evidenciaram a capacidade do modelo em classificar corretamente as instâncias. Para o conjunto de dados mais simples e considerado fácil, obteve-se uma acurácia de 98%, enquanto para os conjuntos médio e difícil, as acurácias obtidas foram de 90% e 83%, respectivamente.

Palavras-chave: Aprendizagem de Máquina, Classificação de Texto, Algoritmo Naïve Bayes.

ABSTRACT

This work consists in a study on modeling a text classifier for automatically categorization of product item descriptions into their appropriate NCM codes. The construction of this classifier was performed using the approach known as supervised learning. The training of this classifier was performed using three sets of data, the first one consisting of NCM transactions from Chapter 20 (Beverages, Liquor and Vinegars), issued by the same retail network; the second composed of transaction items also from Chapter 20, but belonging to different stores; and the third set consisting of transaction items belonging to Chapter 90 (Optical, photographic, cinematographic, medical apparatus and instruments, etc.) and emitted by different establishments. Each of these data sets have different difficulties, were previously labeled and extracted from the database of Consumer Tax Notes (NFC) of the Treasury Department of the State of Rio Grande do Sul.

In addition to that, the classifier training also required that the documents used as input to the classifier were preprocessed text processing techniques, with the goal of transforming them into a suitable format for model calibration. During the preprocessing step, tokenization, stop word removal, stemming and vector extraction of characteristics were performed using TF-IDF measure. At the end of the process, the performance of the model was evaluated using cross-validation methodology with 10-*folds*. The results presented by the proposed approach evidences the ability of the model to correctly classify the instances. For the simples and easier data set, an accuracy of 98% was obtained, while for the medium and difficult sets, the accuracy obtained was 90% and 83%, respectively.

Keywords: Machine Learning, Text Classification, Naïve Bayes Algorithm.

SUMÁRIO

ÍNDICE DE FIGURAS 10
ÍNDICE DE TABELAS 11
ÍNDICE DE ABREVIATURAS 12
ÍNDICE DE ALGORITMOS 13

1. INTRODUÇÃO 14

1.1 PROBLEMA 15
1.2 OBJETIVOS 16
1.2.1 OBJETIVO GERAL 16
1.2.2 OBJETIVOS ESPECÍFICOS 16
1.3 JUSTIFICATIVA 16
1.4 DELIMITAÇÕES DA PESQUISA 17
1.5 ORGANIZAÇÃO DOS CAPÍTULOS 17

2. FUNDAMENTAÇÃO TEÓRICA 19

2.1 NCM – NOMENCLATURA COMUM DO MERCOSUL 19
2.2 ATIVIDADE DE CLASSIFICAÇÃO 21
2.2.1 TEOREMA DE BAYES 23
2.2.2 CLASSIFICADOR NAÏVE BAYES 24
2.3 PROCESSAMENTO DE TEXTO 27
2.3.1 TOKENIZAÇÃO 27
2.3.2 REMOÇÃO DE STOP WORDS 27
2.3.3 STEMMING 28
2.3.4 TERM FREQUENCY (TF) 29
2.3.5 INVERSE DOCUMENT FREQUENCY (IDF) 29
2.3.6 TF-IDF 30
2.4 AVALIAÇÃO DO MODELO 30
2.4.1 VALIDAÇÃO CRUZADA 31
2.4.1.1 MÉTODO *K-FOLD* 31
2.4.1.2 MÉTODO LEAVE-ONE-OUT 32
2.5 CONSIDERAÇÕES 33

3. METODOLOGIA 34

3.1 CARACTERIZAÇÃO DA PESQUISA 34
3.2 PROCEDIMENTOS METODOLÓGICOS 34
3.3 MÉTODO 36
3.4 DEFINIÇÃO DOS EXPERIMENTOS 40
3.4.1 SELEÇÃO DA AMOSTRA 41
3.4.2 AVALIAÇÃO DO MODELO 42
3.5 IMPLEMENTAÇÃO 42
3.6 CONCLUSÕES ..... **Erro! Indicador não definido.**

4. APRESENTAÇÃO DOS RESULTADOS E DISCUSSÕES 45

4.1 EXPERIMENTO SOBRE O CONJUNTO DE DADOS A 45
4.2 EXPERIMENTO SOBRE O CONJUNTO DE DADOS B 45
4.3 EXPERIMENTO SOBRE O CONJUNTO DE DADOS C 46
4.4 COMPARAÇÃO ENTRE RESULTADOS OBTIDOS COM *BASELINE* 47

CONCLUSÃO 49
SUGESTÕES DE TRABALHOS FUTUROS 50
REFERÊNCIAS 52
ANEXO A – Códigos NCM do Capítulo 22 54
ANEXO B – Códigos NCM do Capítulo 90 56

ÍNDICE DE FIGURAS

Figura 1.1 Exemplo de documento de Nota Fiscal Eletrônica ao Consumidor.....14
Figura 2.1 Composição do código NCM.....20
Figura 2.2 - Fluxograma do algoritmo RSLP Stemmer.....29
Figura 2.3 - Exemplo de particionamento de dados do método K-Fold com k=3. 32
Figura 3.1 - Etapas para treinamento e avaliação do classificador.....37

ÍNDICE DE TABELAS

Tabela 1.1 - Trecho do Capítulo 22, Posição 04 da tabela de NCM, destinada a vinhos de uvas frescas, vinhos enriquecidos com álcool e mostos de uvas. 15
Tabela 2.1 – Exemplo da decomposição do NCM 2204.29.11 em Capítulo, Posição, Subposição, Item e Subitem. 21
Tabela 3.1 - Lista de stop words para língua portuguesa. 39
Tabela 3.2 - Número de registros e categorias antes e depois do pré-processamento. 42
Tabela 4.1 - Acurácia média do classificador sobre o conjunto de dados A para os casos com e sem duplicatas. 45
Tabela 4.2 - Acurácia média do classificador sobre o conjunto de dados B para os casos com e sem duplicatas. 46
Tabela 4.3 - Acurácia média do classificador sobre o conjunto de dados C para os casos com e sem duplicatas. 47
Tabela 4.4 - Acurácia média do classificador para os conjuntos de dados A, B e C para os casos com e sem duplicatas. 47
Tabela 4.5 - Acurácia Média dos Capítulos 22 e 90 do SH para o classificador desenvolvido por Ding et al. (2015). 48

ÍNDICE DE ABREVIATURAS

ANSI	*American National Standards Institute*
IDF	*Inverse Document Frequency*
NCM	Nomenclatura Comum do Mercosul
NFC	Nota Fiscal do Consumidor Eletrônica
NLTK	*Natural Language Toolkit*
PLN	Processamento de Linguagem Natural
SEFAZ	Secretaria da Fazenda
SH	Sistema Harmonizado
SQL	*Structured Query Language*
TEC	Tarifa Externa Comum
TF	*Term Frequency*

ÍNDICE DE ALGORITMOS

Algoritmo 1 - Algoritmo para Pré-processamento dos documentos 37
Algoritmo 2 – Algoritmo para treinamento do classificador Naïve Bayes. 38
Algoritmo 3 – Algoritmo para classificação de instâncias não rotuladas..... 39
Algoritmo 4 - Consulta para seleção dos itens pertencentes aos Capítulos 22 e 90 do NCM.. 41

1. INTRODUÇÃO

A criação da versão eletrônica dos documentos fiscais tornou possível a automatização de processos, proporcionando maior agilidade e confiabilidade das atividades contábeis e fiscais. A Figura 1.1 exibe um documento de Nota Fiscal Eletrônica ao Consumidor (NFC-e). Esse documento é a versão impressa do documento digital que está disponível no site da Receita Estadual, podendo o acesso a esse ser realizando através do site http://www.sefaz.rs.gov.br/NFCE, mediante a utilização da chave de acesso informada no corpo do documento.

Figura 1.1 Exemplo de documento de Nota Fiscal Eletrônica ao Consumidor

DIMED S/A DISTRIBUIDORA DE MEDICAMENTOS F. 101 POA

CNPJ: 92.665.611/0134-06 Inscrição Estadual: 0962203505

PRAIA DE BELAS, 1181, PRAIA DE BELAS, PORTO ALEGRE, RS

DANFE NFC-e - Documento Auxiliar da Nota Fiscal Eletrônica para Consumidor Final - Via Consumidor

NFC-e não permite aproveitamento de crédito de ICMS

Emissão normal

NFC-e nº: 17690 Série: 1 Data de Emissão: 02/07/2015 12:10:29

Consulte pela Chave de Acesso em https://www.sefaz.rs.gov.br/NFCE

CHAVE DE ACESSO

4315 0792 6656 1101 3406 6500 1000 0176 9013 7724 5110

Protocolo de Autorização: 143150042457601

CONSUMIDOR

CPF: 007.161.950-03 - RODRIGO DE ABREU BATISTA

Código	Descrição	Qtde	Un	Vl Unit	Vl Total
7081	TANDRILAX 30 CP	1	CX	24	24,00

Valor total R$	22,50
Valor descontos R$	1,50
FORMA PAGAMENTO	VALOR PAGO R$
Cartão de Débito	22,50

Versão XSLT: 1.10

Fonte: SEFAZ RS (https://www.sefaz.rs.gov.br/NFE/NFE-NFC.aspx)

Através do acesso à versão digital desse documento é possível obter informações detalhadas da nota e itens de produtos, que são omitidas na versão impressa. Entre essas informações estão o código NCM dos itens e as alíquotas de ICMS aplicadas aos mesmos. A sigla NCM é um acrônimo para Nomenclatura Comum do Mercosul, e trata-se de um código hierárquico de categorias através dos quais são classificados os produtos. A cada código de NCM estão atreladas descrições e alíquotas de ICMS. Por exemplo, a Tabela 1.1 exibe as informações NCM, Descrição e TEC (Tarifa Externa Comum) e corresponde a parte do Capítulo 22, Posição 04 da tabela de NCM, correspondente a vinhos de uvas frescas, vinhos

enriquecidos com álcool e mostos de uvas, extraída do site da Receita Estadual.

Fica a critério de cada estabelecimento comercial informar a descrição de cada produto que aparecerá na Nota Fiscal do Consumidor (NFC-e). Também é de responsabilidade do estabelecimento vincular a cada um desses produtos descritos o correspondente código NCM. Cada NCM possui um valor de alíquota atrelado, correspondente ao imposto que incidirá sobre a venda do produto. Alguns contribuintes, agindo por engano ou intencionalmente, associam códigos NCM que não correspondem, de fato, aos itens de produtos descritos, de modo que alíquotas diferentes passem a incidir sobre os mesmos. Dessa forma, o estado deixaria de estar arrecadando o referido imposto, sendo que em uma situação como essa, o contribuinte envolvido deveria ser autuado.

Tabela 1.1 - Trecho do Capítulo 22, Posição 04 da tabela de NCM, destinada a vinhos de uvas frescas, vinhos enriquecidos com álcool e mostos de uvas.

NCM	DESCRIÇÃO	TEC (%)
22.04	**Vinhos de uvas frescas, incluindo os vinhos enriquecidos com álcool; mostos de uvas, excluindo os da posição 20.09.**	
2204.10	- Vinhos espumantes e vinhos espumosos	
2204.10.10	Tipo champanha (*champagne*)	20
2204.10.90	Outros	20
2204.2	- Outros vinhos; mostos de uvas cuja fermentação tenha sido impedida ou interrompida por adição de álcool:	
2204.21.00	-- Em recipientes de capacidade não superior a 2 l	20
2204.29	-- Outros	
2204.29.1	Vinhos	
2204.29.11	Em recipientes de capacidade não superior a 5 l	20
2204.29.19	Outros	20
2204.29.20	Mostos	20
2204.30.00	- Outros mostos de uvas	20

Fonte: http://www.mdic.gov.br/comercio-exterior/estatisticas-de-comercio-exterior-9/

1.1 PROBLEMA

As divergências existentes entre descrições de produtos e códigos NCM informados pelos estabelecimentos em Notas Fiscais do Consumidor (NFC-e), consistem no problema que originou esse trabalho. Dessa forma, o presente trabalho busca responder a questão relacionada a como classificar automaticamente os itens de NFC de modo a tornar possível a identificação dessas divergências.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

O objetivo geral desse trabalho é classificar os itens de Nota Fiscal Eletrônica do Consumidor (NFC-e) com base nas descrições textuais dos produtos, atribuindo-os aos códigos NCM correspondentes. Para isso, desenvolveu-se um classificador que com base na descrição do produto, informa o código NCM ao qual o mesmo deve pertencer. Esse classificador foi testado sobre registros de transações obtidos a partir do banco de dados da Secretaria da Fazenda do Rio Grande do Sul (SEFAZ-RS).

1.2.2 OBJETIVOS ESPECÍFICOS

Ao longo do desenvolvimento desse trabalho, atingiu-se os seguintes objetivos específicos:

- Pesquisar técnicas eficientes de processamento de texto que possibilitem lidar com grandes volumes de dados de modo a torná-los adequados para o treinamento de classificadores.
- Estudar técnicas de classificação de texto que sejam escaláveis e eficazes para a utilização de classificadores online.
- Implementar um classificador de texto que classifique descrições textuais de itens de produtos, associando-os a códigos NCM adequados.
- Analisar os resultados do classificador implementado utilizando métricas adequadas para avaliação de modelos, a fim de validar o desempenho do classificador treinado.
- Testar o classificador sobre dados reais de transações obtidas a partir do banco de dados da SEFAZ-RS.

1.3 JUSTIFICATIVA

A motivação para o desenvolvimento desse trabalho surgiu da necessidade de encontrar divergências entre descrições de produtos e os respectivos códigos NCM

informados em NFC-e durante o desenvolvimento da atividade de auditoria realizada pela SEFAZ-RS. Essas divergências podem ser originadas tanto pela falta de informação dos contribuintes sobre a classificação NCM, como pela tentativa de diminuir as alíquotas incidentes sobre os produtos vendidos, com o objetivo de recolher menos imposto.

Dessa forma, a contribuição desse trabalho se dá essencialmente em duas frentes. A primeira delas é no campo tecnológico, ao abordar o tema classificação, que é uma das atividades de aprendizagem de máquina (*machine learning*), aplicado ao setor público, onde, no Brasil, recém estão surgindo as primeiras aplicações dessa tecnologia. A segunda contribuição ocorre no campo socioeconômico, uma vez que ao munir o governo com mecanismos eficientes no combate à corrupção, também está se auxiliando no aumento da arrecadação e, consequentemente, no aumento dos investimentos e benefícios para a população.

1.4 DELIMITAÇÕES DA PESQUISA

Esta pesquisa constitui-se no estudo de técnicas e ferramentas necessárias para o desenvolvimento de um classificador de descrições de produtos em códigos NCM. Para isso, está compreendido o estudo de técnicas de processamento de texto e aprendizagem de máquina. Espera-se que a pesquisa mobilize a criação do protótipo funcional de um classificador que possa, futuramente, ser utilizado largamente para contribuir na atividade de auditoria eletrônica ligada à SEFAZ-RS. Não está previsto nesse trabalho o desenvolvimento de um uma solução completa que compreenda todos os códigos NCM previstos atualmente pela Secretaria da Fazenda. Em vez disso, optou-se por utilizar apenas um subconjunto dos códigos NCM existentes para validação da proposta, permitindo assim que o foco do trabalho se mantivesse no método escolhido, e não em questões decorrentes da problemática envolvida ao lidar-se com grandes volumes de dados.

1.5 ORGANIZAÇÃO DOS CAPÍTULOS

Esse trabalho está organizado da seguinte maneira:

O Capítulo 1 introduz o tema abordado ao longo do trabalho, a motivação, os objetivos e a organização. O Capítulo 2 discute os conceitos relacionados à Nomenclatura Comum do Mercosul, NCM, aprendizado Bayesiano e processamento de texto. O Capítulo 3 apresenta a

abordagem proposta, detalhando como é realizado o processamento de texto e treinamento do classificador. O Capítulo 4 descreve os experimentos realizados e a análise de desempenho do classificador desenvolvido, além da discussão dos resultados obtidos. Por fim, a Conclusão apresenta considerações sobre os resultados e a contribuição desse trabalho, além de sugerir possíveis extensões que podem vir a ser abordadas como trabalhos futuros. O Anexo A apresenta a tabela completa de códigos e descrições de NCMs pertencentes ao Capítulo 22 e, o Anexo B, apresenta a tabela completa de códigos e descrições de NCMs pertencentes ao Capítulo 90.

2. FUNDAMENTAÇÃO TEÓRICA

O objetivo desse capítulo é introduzir os conceitos e técnicas básicas que serão utilizadas ao longo do texto. A Seção 2.1 apresenta a Nomenclatura Comum do Mercosul (NCM), assim como um breve histórico e importância. A Seção 2.2 descreve e formaliza a atividade de classificação, o teorema de Bayes e o classificador Naïve Bayes. A Seção 2.3 introduz os conceitos e técnicas envolvidas na atividade de processamento de texto que são utilizados no pré-processamento dos documentos pelo classificador. A Seção 2.4 aborda as métricas básicas utilizadas na avaliação de modelos, juntamente da metodologia de validação cruzada. Por fim, a Seção **Erro! Fonte de referência não encontrada.** apresenta as conclusões relacionadas aos tópicos apresentados nesse capítulo.

2.1 NCM – NOMENCLATURA COMUM DO MERCOSUL

A Nomenclatura Comum do Mercosul, mais comumente conhecida pelo acrônimo NCM, é uma convenção de categorização de mercadorias adotada desde 1995 pelos países que compõem o bloco econômico do Mercosul, sendo esses Uruguai, Paraguai, Brasil e Argentina. O NCM toma como base e estende o Sistema Harmonizado (SH), criado em 1983 pela Organização Mundial das Alfândegas (OMA) e utilizado internacionalmente para padronizar e classificar produtos de importação e exportação (RECEITA FEDERAL, 2015).

Antes da adoção desses sistemas ficava a critério de cada organização descrever textualmente, e a seu modo, cada uma de suas mercadorias, o que poderia gerar inconsistências e ambiguidades, além de esbarrar em questões idiomáticas quando em nível internacional. A padronização de mercadorias através de códigos proporcionou a facilitação de negociações comerciais, além de simplificar o levantamento de estatísticas internacionais. Isso só foi possível, pois uma mesma mercadoria passou a ser conhecida por um único código internacional em nível mundial. Em termos práticos, o NCM vem sendo utilizado na elaboração de tarifas de direitos aduaneiros e de frete, de estatísticas de comércio de importação e de exportação, de produção e dos diferentes meios de transporte de mercadorias.

Os códigos do NCM são compostos por oito dígitos, sendo os seis primeiros dígitos herdados do Sistema Harmonizado, acrescido de dois dígitos, que são específicos do âmbito do Mercosul. A **Erro! Fonte de referência não encontrada.** ilustra como é composta essa codificação. Os seis primeiros dígitos correspondem a três pares de dígitos que especificam o

Capítulo, a Posição e a Subposição, todos esses originados do Sistema Harmonizado. O sétimo e oitavo dígitos existem exclusivamente na codificação NCM, e referem-se ao Item e Subitem, respectivamente.

Figura 2.1 Composição do código NCM.

Fonte: Receita do Estado do RS.

Tomando-se como exemplo o NCM de código 2204.29.11, seus dígitos podem ser decompostos conforme a Tabela 2.1, e refere-se a Vinhos armazenados em recipientes de capacidade não superior a 5 litros. Nesse caso, os dois primeiros dígitos referem-se ao Capítulo 22, que diz respeito a “Bebidas, líquidos alcoólicos e vinagres”. O terceiro e quarto dígitos referem-se à Posição 04, que diz respeito a “Vinhos de uvas frescas, incluindo os vinhos enriquecidos com álcool; mostos de uvas, excluindo os da posição 20.09.”. O quinto e sexto dígitos referem-se à Subposição 29, que na tabela de NCM consta como “--Outros”. O sétimo dígito refere-se ao Item 1, que corresponde a “Vinhos”, enquanto que o Subitem 1 diz respeito a “recipientes de capacidade não superior a 5 l”. As tabelas de NCM referentes aos Capítulos 22 e 90, que serão utilizadas ao longo desse texto como formas de ilustração, estão listadas nos Anexos A e B, respectivamente.

O NCM está estruturado em 99 Capítulos, que dizem respeito a categorias gerais, e que estão agrupados em 21 seções. Os Capítulos estão organizados de acordo com um ordenamento lógico, em ordem crescente de sofisticação ou participação humana na produção do bem. Dessa forma, o primeiro capítulo refere-se à categoria Animais Vivos, enquanto que o último refere-se a Obras de Arte. O Capítulo 77 permanece em branco e está destinado a uma eventual utilização futura pelo Sistema Harmonizado. Os Capítulos 98 e 99 constam como reservados para usos especiais pelas partes contratantes. No Brasil, o Capítulo 99 é utilizado para registrar operações especiais na atividade de exportação (MINISTÉRIO DA INDÚSTRIA, COMÉRCIO EXTERIOR E SERVIÇOS, 2016).

Tabela 2.1 – Exemplo da decomposição do NCM 2204.29.11 em Capítulo, Posição, Subposição, Item e Subitem.

Componente	**Valor**	**Descrição**
Capítulo	22	Bebidas, líquidos alcoólicos e vinagres
Posição	04	Vinhos de uvas frescas, incluindo os vinhos enriquecidos com álcool; mostos de uvas, excluindo os da posição 20.09.
Subposição	29	-- Outros
Item	1	Vinhos
Subitem	1	Em recipientes de capacidade não superior a 5 l

Fonte: Autor.

2.2 ATIVIDADE DE CLASSIFICAÇÃO

A classificação é uma das atividades de mineração de dados que consiste em associar rótulos de classe a um conjunto de instâncias (casos) não classificadas. Essa classificação pode ser de dois tipos, podendo ser supervisionada e não supervisionada. Na classificação supervisionada, o conjunto de possíveis classes é conhecido previamente, enquanto que na classificação não supervisionada, o conjunto de classes possíveis não é conhecido. Após um processo de classificação não supervisionada, é provável que se deseje atribuir nomes às classes (RUSSEL e NORVIG, 2003). A classificação não supervisionada é também chamada de agrupamento (*clustering*), e consiste em agrupar itens similares com base em uma função de distância. Ao longo do desenvolvimento desse trabalho será abordada a classificação do tipo supervisionada.

A atividade de classificação de textos é atualmente realizada em diversas áreas que manipulam grandes volumes de texto, que vão desde agências de notícias a motores de busca. Entre os exemplos dessa atividade estão a identificação de linguagem (classes: Inglês vs. Português), detecção automática de spam em e-mails (classes: spam vs. Não spam), detecção de conteúdo erótico (erótico vs. não-erótico), detecção de sentimento em relação a opiniões baseadas em filmes e avaliação de produtos (classes: positivo vs. negativo).

Inicialmente, essa atividade era realizada manualmente por especialistas. A empresa de buscas Yahoo, no início da web, utilizava classificação manual para organizar as notícias em seu portal. Essa atividade era extremamente precisa e consistente, quando o conjunto de documentos a ser classificado e a equipe permanecem pequenos. No entanto, alcançar a

escalabilidade nesse tipo de atividade, torna-se difícil e caro, exigindo, assim, métodos automáticos de classificação (MANNING *et al.*, 2008).

O passo seguinte veio através da classificação de notícias através de regras de classificação. O Google Alerts, serviço que vende a facilidade de monitorar a web em busca de conteúdo novo e interessante, era baseado nessa tecnologia, e utilizava regras montadas através de combinações booleanas. Apesar de possuir uma acurácia alta, construir e manter sistemas baseados em regras torna-se trabalhoso e muito caro, a medida em que se aumenta a quantidade de regras, pois essas devem ser cuidadosamente refinadas ao longo do tempo por especialistas (MANNING *et al.*, 2008).

O passo seguinte na evolução dos classificadores passou a utilizar métodos de classificação estatísticos e probabilísticos. Tais métodos baseiam-se em utilizar dados previamente rotulados para a criação de modelos que serão aplicados para classificar novos dados. Esse tipo de estratégia ainda requer que sejam rotuladas instâncias de treinamento que serão utilizadas para ajustar um modelo, mas possuem a capacidade suficiente de generalização para que possa ser utilizado para classificar novas instâncias, livrando o especialista da tarefa custosa de classificação manual e permitindo assim lidar com grades volumes de dados (MANNING *et al.*, 2008).

Atualmente, a utilização de métodos estatísticos e probabilísticos, combinada a técnicas desenvolvidas no ramo da ciência da computação como redes neurais, programação dinâmica, inteligência artificial e aprendizado profundo (*deep learning*), deram origem a uma subárea da ciência da computação, chamada aprendizagem de máquina (*machine learning*). Mitchell (1997), em seu livro Machine Learning, define aprendizagem de máquina da seguinte forma: “Se diz que um programa de computador aprende pela experiência E em relação a algum tipo de tarefa T com performance P, se sua performance P nas tarefas em T melhoram com a experiência E”.

Nesse trabalho será utilizada a abordagem de classificação supervisionada de documentos. O problema de classificação pode ser formalizado como segue:

Dados:

- Espaço de documentos X: os documentos são representados nesse espaço – tipicamente, algum espaço com grande número de dimensões.
- Conjunto finito de classes $C = \{c_1, c_2, \dots, c_m\}$: as classes são definidas manualmente, de acordo com as necessidades da aplicação (ex. relevante vs.

não-relevante).

- Conjunto de treinamento D de documentos classificados, com cada documento classificado $\langle d, c \rangle \in X \times C$.

Usando o Algoritmo Naïve Bayes, deseja-se construir um classificador Υ que seja capaz de mapear documentos em classes $\Upsilon: X \rightarrow C$. A aplicação do modelo treinado Υ a uma descrição $d \in X$ de um documento dá-se através da determinação de $\Upsilon(d) \in C$, que representa a classe mais apropriada para d.

2.2.1 TEOREMA DE BAYES

O conceito de probabilidade condicional, introduzido pela estatística elementar, define que a probabilidade condicional de um evento é a probabilidade obtida pela informação adicional de algum outro evento que já ocorreu. A Equação 1 pode ser utilizada para encontrar $P(A|B)$, ou seja, a probabilidade de A dada a ocorrência de B.

$$P(A|B) = \frac{P(A\ e\ B)}{P(B)} \tag{1}$$

Na teoria da probabilidade, o Teorema de Bayes mostra a relação entre uma probabilidade condicional e sua inversa, isso é, a probabilidade de uma hipótese ser verdadeira dada a observação de uma evidência, e a probabilidade da evidência dada pela hipótese (TRIOLA, 2008). O Teorema de Bayes representa uma das primeiras tentativas de se modelar matematicamente a inferência estatística. Dessa forma, a probabilidade de ocorrência de um evento A, dada a observação de B, pode ser obtida pelo cálculo de probabilidade representado na Equação 2, onde $P(A)$ e $P(B)$ são as probabilidades *a priori* de A e B, e $P(B|A)$ e $P(A|B)$ são as probabilidades *a posteriori* de B condicionada a A e de A condicionada a B, respectivamente. Probabilidade *a priori* é a probabilidade inicial obtida antes de qualquer informação adicional ser fornecida, enquanto que probabilidade *a posteriori* refere-se ao valor da probabilidade atualizado, utilizando-se a informação adicional fornecida pela evidência.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (2)$$

2.2.2 CLASSIFICADOR NAÏVE BAYES

Em aprendizagem de máquina, classificadores Naïve Bayes são uma família de classificadores probabilísticos baseados na aplicação do Teorema de Bayes. O classificador leva em seu nome o adjetivo *naïve* (ingênuo), por apresentar a forte premissa de existir independência entre as características (*features*) das instâncias a serem classificadas. Classificadores Naïve Bayes são altamente escaláveis e, por isso, adequados para serem utilizados em aplicações online e de tempo real. O treinamento de um modelo é feito através da avaliação de uma expressão fechada, que possui tempo linear, em vez de aproximações iterativas e custosas utilizadas por outros tipos de classificadores (RUSSEL e NORVIG, 2003).

Naïve Bayes é uma técnica simples e eficaz para construir classificadores. Classificadores são modelos que associam rótulos a instâncias-problema, representadas por vetores de características. Um vetor de características é um vetor $x = (x_1, \dots, x_n)$ que representa as n características (variáveis independentes) de cada uma das instâncias do conjunto a ser classificado. Não existe um único algoritmo para treinar esses classificadores, mas uma família de algoritmos que podem ser utilizados. Todos esses algoritmos assumem como premissa o fato de todas as variáveis serem independentes, isso é, o valor de uma característica não depende de qualquer outra característica, dada qualquer classe (MANNING *et al.*, 2008). Por exemplo, uma fruta pode ser considerada uma maçã se ela é vermelha, redonda e possui cerca de 10 cm de diâmetro. Um classificador Naïve Bayes considera que cada uma dessas características contribui independentemente para a probabilidade dessa fruta ser uma maçã, desconsiderando qualquer possível correlação existente entre as características cor, formato e diâmetro.

No que diz respeito ao domínio das variáveis presentes no vetor de características, diferentes variantes do classificador Naïve Bayes podem ser utilizadas. Entre esses modelos, estão os seguintes (MANNING *et al.*, 2008):

- Naïve Bayes Gaussiano: utilizado para lidar com variáveis contínuas. Tem como premissa de que a distribuição dos valores de suas variáveis segue a distribuição normal.

- Naïve Bayes Multinomial: utilizado para lidar com variáveis discretas.
- Naïve Bayes Bernoulli: utilizado para lidar com variáveis booleanas.

Naïve Bayes tem sido estudado extensivamente desde a década de 50, tendo sido introduzido sob um nome diferente na comunidade de recuperação de informação no início dos anos 60, permanecendo um método popular e usado como *baseline* para a classificação de textos (RUSSEL e NORVIG, 2003). O problema da classificação de texto consiste em categorizar documentos como pertencendo a uma categoria ou outra, utilizando a frequência de ocorrência das palavras como características. A frequência de ocorrência das palavras é realizada em uma etapa prévia à classificação, e deve ser feita para cada texto a ser categorizado.

A probabilidade de um documento d pertencer à classe c pode ser computada conforme a Equação 3, onde:

- n_d: refere-se ao comprimento do documento (número de *tokens*);
- $P(t_k|c)$: refere-se à probabilidade condicional do termo t_k ocorrer em um documento da classe c. Também pode ser visto como uma medida de quanto a presença de t_k contribui para determinar que c é a classe correta do documento;
- $P(c)$: refere-se à probabilidade a priori de c.

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \tag{3}$$

Além disso, caso os termos de um documento não forneçam evidências suficientes para determinar a qual classe o documento pertence, atribui-se a classe c com maior $P(c)$. Como o objetivo com o algoritmo Naïve Bayes é encontrar a melhor classe para o documento, muito provavelmente deseja-se obter a classe máxima a posteriori c_{map}, ou seja, a classe com maior probabilidade quando as probabilidades condicionais são consideradas. Dessa forma, a obtenção de c_{map} pode ser obtida conforme a Equação 4.

$$c_{map} = arg \max_{c \in C} \hat{P}\,(c|d) = arg \max_{c \in C} \hat{P}\,(c) \sum_{1 \leq k \leq n_d} \hat{P}(t_k|c) \tag{4}$$

A multiplicação de um número grande de probabilidades pode resultar em um número muito pequeno, ocasionando em um possível *underflow* de ponto flutuante. Uma maneira de contornar isso é através do uso de logaritmos. Utilizando-se da propriedade de logaritmos expressa na Equação 5 e, sendo log uma função monotônica, pode-se utilizar desse artifício para a identificação da classe com maior escore, uma vez que a classe com maior probabilidade não mudará após a aplicação do logaritmo. A equação modificada para cálculo de c_{map} após a aplicação do logaritmo está descrita na Equação 6.

$$\log(xy) = log(x) + \log(y) \tag{5}$$

$$c_{map} = arg \max_{c \in C} \left[\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c) \right] \tag{6}$$

Na Equação 6, a contribuição de cada parâmetro condicional, representado por $\log \hat{P}(t_k|c)$, reflete o quão bom é para o indicador a presença do *token* t_k para a classe c. Já a probabilidade *a priori*, representada na equação pela componente $\log \hat{P}(c)$, é uma medida que reflete exclusivamente a frequência relativa de c. Dessa forma, a soma da probabilidade a priori com os pesos dos termos é uma medida de quanta evidência existe para afirmar que um documento pertence a uma determinada classe. Ao final do processo, $arg \max_{c \in C}$ refere-se à classe com maior evidência c_{map}. Os parâmetros $\hat{P}(c)$ e $\hat{P}(t|c)$ podem ser estimados a partir das Equações 7 e 8, respectivamente, onde:

- N: refere-se ao número total de documentos;
- N_c: refere-se ao número de documentos na classe c;
- $T_{c,t}$: refere-se ao número de ocorrências de *tokens* t nos documentos de treinamento da classe c.

$$\hat{P}(c) = \frac{N_c}{N} \tag{7}$$

$$\hat{P}(t|c) = \frac{T_{c,t}}{\sum_{t' \in V} T_{ct'}} \tag{8}$$

2.3 PROCESSAMENTO DE TEXTO

A atividade de processamento de texto consiste na tarefa de converter texto puro armazenado em um arquivo, essencialmente uma sequência digital de bits, em uma sequência bem definida de unidades linguisticamente significativas: no nível mais baixo, caracteres representando os morfemas do sistema de escrita de uma linguagem; palavras, formadas consistindo de um ou mais caracteres; e sentenças, formadas por uma ou mais palavras (INDURKHYA e DAMERAU, 2010). O processamento de texto é uma parte fundamental de um sistema de processamento de linguagem natural (PLN), visto que todo o tipo de análise realizada sobre um texto depende da identificação dos caracteres, palavras e sentenças realizadas nessa etapa.

2.3.1 TOKENIZAÇÃO

A segmentação de texto é o processo de converter um texto em sentenças e palavras. A segmentação por palavra é realizada pela da quebra da sequência de caracteres através da localização dos limites de uma palavra, ou seja, os pontos de início e fim da palavra. Para fins linguísticos e computacionais, as palavras identificadas desse modo são chamadas de *tokens*, e o processo de segmentação de palavras é chamado de tokenização (INDURKHYA e DAMERAU, 2010).

2.3.2 REMOÇÃO DE STOP WORDS

Em processamento de texto, *stop words*, ou palavras de parada, referem-se às palavras mais comuns de uma linguagem. Tais palavras são frequentemente removidas antes ou depois do processamento de texto, por pouco contribuírem, por exemplo, em análises ou classificações.

Não existe uma lista universal de *stop words* usada por todas as ferramentas de processamento de linguagem natural. Em vez disso, qualquer grupo de palavras pode ser escolhido de acordo com o propósito desejado. Frequentemente as listas de *stop words* são criadas tendo como base análises estatísticas sobre os textos que pertencem ao conjunto de textos a serem analisados (*corpus*) (LESKOVEC *et al.*, 2014). Motores de busca

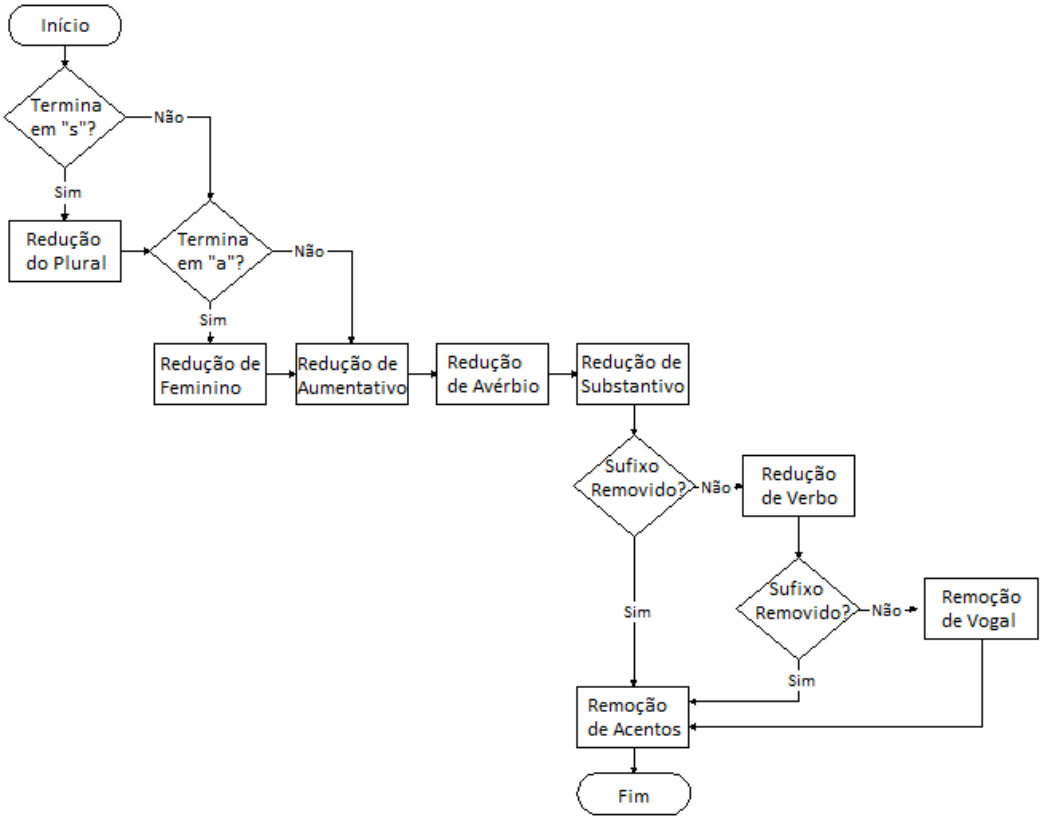
frequentemente incluem em suas listas de *stop words* as preposições e os artigos de uma linguagem, por serem termos frequentemente comuns e de pouca contribuição para análises (MANNING *et al.*, 2008).

2.3.3 STEMMING

Stemming é o nome dado ao processo de juntar as formas variantes de uma palavra em uma expressão comum, o radical. Radical é o elemento que contém o significado básico da palavra, sendo a ele acrescentados outros elementos. A técnica de *stemming* consiste basicamente em remover os sufixos das palavras. Esse processo é largamente usado em recuperação de informações com o objetivo de aumentar a revocação, ou seja, a quantidade de documentos retornada a partir de uma palavra chave utilizada como busca (ORENGO e HUYCK, 2001). Por exemplo, as palavras pedra, pedreiro, pedregulho, pedrada, apedrejar e pedreira possuem o mesmo radical, que é “ped-“, possibilitando que uma consulta realizada utilizando um desses termos retornasse documentos que possuam outras palavras derivadas do mesmo radical.

A Figura 2.2 exibe o fluxograma do algoritmo RSLP, um algoritmo de *stemming* para língua Portuguesa desenvolvida por Orengo e Hyuck (2001). Nessa figura está ilustrada a sequência de passos a serem executados para obtenção do radical de uma palavra. Cada uma dessas etapas consiste na avaliação de um conjunto de regras, das quais somente uma será aplicada. Ao final de cada etapa, o maior sufixo possível será removido.

Figura 2.2 - Fluxograma do algoritmo RSLP *Stemmer*.



Fonte: Adaptado de Orengo e Huyck (2001).

2.3.4 TERM FREQUENCY (TF)

A medida TF (*Term Frequency*) corresponde à frequência com que um termo ocorre em um documento. A frequência de um termo t em um documento d, denotada por $TF(t,d)$, corresponde ao número de vezes que t ocorre em d. A medida TF normalmente é representada em número absoluto, podendo ser usada uma variação booleana, que atribui $TF(t,d) = 1$, caso t ocorra em d e $TF(t,d) = 0$, caso contrário (LUHN, 1957).

2.3.5 INVERSE DOCUMENT FREQUENCY (IDF)

A medida IDF (*Inverse Document Frequency*), corresponde ao inverso da frequência com que um dado termo ocorre nos documentos. Essa medida reflete quanto de informação um dado termo fornece, isso é, se o termo é comum ou raro considerando-se a totalidade do conjunto de documentos (SPARCK JONES, 1972). A Equação 9 representa como se dá o cálculo da medida IDF, onde N refere-se ao número total de documentos no *corpus* ($N =$

$|D|$) e $|\{d \in D: t \in d\}|$ refere-se ao número de documentos onde o termo t aparece.

$$IDF(t, D) = \log \frac{N}{|\{d \in D: t \in d\}|} \tag{9}$$

2.3.6 TF-IDF

A métrica TF-IDF, um acrônimo para *Term Frequency-Inverse Document Frequency*, é uma medida estatística que tem como objetivo refletir o grau de importância de uma palavra em um documento, levando em consideração toda a coleção de documentos (*corpus*) sob análise. Essa medida é frequentemente utilizada como fator de ponderação para utilização em algoritmos de recuperação de informações, classificação e mineração de texto (LESKOVEC, 2014).

O cálculo de TF-IDF se dá através do produto entre as métricas TF e IDF, conforme expresso pela Equação 10. Dessa forma, altos valores de TF-IDF serão obtidos com para termos com alta frequência em um dado documento e com baixa frequência sobre a coleção de documentos. Por outro lado, TF-IDF produzirá baixos valores para termos presentes em muitos documentos. Essas características permitem que essa métrica seja adequada na identificação de termos significativos e na filtragem de termos relativamente comuns.

$$TFIDF(t, d, D) = TF(t, d).IDF(t, D) \tag{10}$$

2.4 AVALIAÇÃO DO MODELO

O treinamento de um modelo e teste utilizando os mesmos dados é um erro comum de metodologia, pois ainda que um modelo consiga reproduzir os rótulos das instâncias recém-treinadas com boa performance, ele pode falhar ao classificar dados úteis ainda não vistos. Essa situação é chamada *overfitting*, e ocorre quando um modelo se torna tão adaptado aos dados que falha ao classificar novas instâncias. No aprendizado supervisionado, uma prática comum para evitar o *overfitting* é separar uma parte dos dados disponíveis para ser utilizada exclusivamente como conjunto de teste.

Esse método, conhecido por *holdout*, consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treinamento (calibragem do modelo) e

outro para teste (validação). O conjunto de dados pode ser separado em quantidades iguais ou não. Uma proporção comum é considerar 2/3 dos dados para treinamento e o 1/3 para teste (KOHAVI, 1995).

Após o particionamento, o treinamento do modelo é realizado e, posteriormente, os dados de teste são apresentados para avaliação do desempenho do classificador. Esta abordagem é indicada quando se tem à disposição uma grande quantidade de dados. Caso o conjunto total de dados seja pequeno, o erro gerado na rotulação pode sofrer muita variação. Quando se dispõe de um conjunto de dados reduzido, uma maneira apropriada de estimar o desempenho preditivo de um modelo é através de validação cruzada.

2.4.1 VALIDAÇÃO CRUZADA

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo a partir de um conjunto de dados reduzido (KOHAVI, 1995). Esse método se baseia na divisão do conjunto total de dados em outros dois subconjuntos mutuamente exclusivos, sendo o primeiro desses destinado a treino e o segundo destinado a teste. Com o objetivo de reduzir a variabilidade, esse processo é repetido iterativamente por diversas vezes, selecionando-se a cada repetição diferentes instâncias para compor os conjuntos de treinamento e teste.

Outra das principais razões para usar validação cruzada no lugar da validação convencional (por exemplo, particionar os dados em dois conjuntos, 70% para treinamento e 30% para teste) é a insuficiência de dados disponíveis para particionar o conjunto em dois subconjuntos sem a perda da capacidade de modelagem ou teste. Para isso, a validação cruzada promove a rotatividade dos dados entre os conjuntos, permitindo que todas as instâncias participem da fase de treinamento e teste.

2.4.1.1 MÉTODO *K-FOLD*

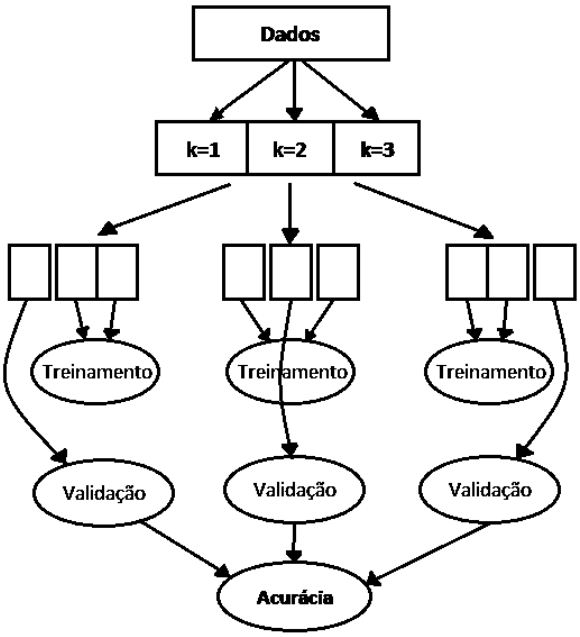
O método de validação cruzada denominado *K-Fold* consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos e de mesmo tamanho (*fold*). O primeiro desses subconjuntos é então separado para teste, enquanto que os demais $k - 1$ subconjuntos são utilizados para treinamento do modelo. Ao final de cada etapa, a taxa de erro do classificador treinado pode ser calculada através da Equação 11. O processo de

treinamento e validação do modelo é realizado por k vezes, alternando-se de forma circular o subconjunto de teste, conforme representado na Figura 2.3. Ao final das k iterações, calcula-se a acurácia média sobre os resultados encontrados, através da Equação 12, obtendo assim uma medida mais confiável sobre a capacidade do modelo de classificar corretamente os dados.

$$Err(\Upsilon) = \frac{1}{n}\sum_{1}^{n} \|y_i \neq \Upsilon(x_i)\| \tag{11}$$

$$Acc(\Upsilon) = 1 - Err(\Upsilon) \tag{12}$$

Figura 2.3 - Exemplo de particionamento de dados do método K-Fold com k=3.



Fonte: adaptado de https://pt.wikipedia.org/wiki/Validação_cruzada

2.4.1.2 MÉTODO LEAVE-ONE-OUT

O método *leave-one-out* é um caso específico do método *K-Fold*, onde k é igual ao número total de dados N. Nesta abordagem são treinados N modelos, cada um desses avaliados sobre uma única instância. Esse método possui a vantagem de ser uma investigação completa sobre a variação do modelo em relação aos dados utilizados, no entanto, possui alto

custo computacional associado, sendo indicado em situações onde poucos dados estão disponíveis.

2.5 CONSIDERAÇÕES

Nesse capítulo foi apresentada a Nomenclatura Comum do Mercosul (NCM) e ressaltada sua importância para as transações nacionais e internacionais. Além disso, detalhou-se sua estrutura de característica hierárquica e foram traçados os pontos em comum com o Sistema Harmonizado (SH), no qual o NCM é baseado.

Nesse capítulo também foram introduzidos conceitos relacionados à classificação de dados, apresentado o Teorema de Bayes e descrito o classificador Naïve Bayes, que trata-se de uma simplificação do classificador Bayesiano, por assumir que existe a independência entre as variáveis. Apesar de existirem outros métodos de classificação, optou-se por detalhar a abordagem bayesiana e o algoritmo Naïve Bayes, por ser uma técnica escalável e com performance adequada para ser utilizada em aplicações online e de tempo real, que exijam rápida resposta, devido à sua característica de resolução das categorias através de uma fórmula fechada, em oposição a outros métodos iterativos e de maior complexidade.

Na Seção 2.3 foram abordados os conceitos e técnicas relacionadas ao processamento de texto. Ao longo da seção foram introduzidos conceitos como tokenização, remoção de *stop words* e *stemming*, além das métricas *Term Frquency* (TF), *Inverse Document Frequency* (IDF) e TF-IDF. A combinação dessas técnicas foi utilizada no pré-processamento dos documentos de NFC, de modo a transformar os textos em vetores de características (*features*), utilizados na calibragem do modelo de classificação. Por fim, a Seção 2.4 foi destinada à avaliação de modelo e descreve a importância em se adotar uma metodologia de validação. As métricas de desempenho baseadas em erro e acurácia, além do método de validação cruzada *K-Fold*, também foram abordadas nessa seção. A medida de acurácia foi escolhida para avaliar o modelo, pois reflete, em termos práticos, o que se espera de um modelo de classificação eficiente, ao representar a proporção entre a quantidade de classificações corretas em relação ao total de classificações realizada. O método K-*fold* foi escolhido por ser uma metodologia de avaliação que busca eliminar possíveis vieses existentes na seleção dos dados destinados a treinamento e teste, fornecendo valores de acurácia mais precisos.

3. METODOLOGIA

Este capítulo tem como objetivo apresentar o método utilizado para classificação de descrições de itens de NFC em códigos de NCM correspondentes. O método consiste em treinar um classificador Naïve Bayes utilizando-se de instâncias previamente rotuladas, e avaliar o desempenho do mesmo através da validação sobre um conjunto de instâncias de teste.

3.1 CARACTERIZAÇÃO DA PESQUISA

A pesquisa se caracteriza como empírica e exploratória. Empírica por utilizar-se de métodos estatísticos e metodologia aceita pela comunidade para comparação do método proposto e, exploratória, pois busca aplicar e analisar os resultados obtidos, na utilização de técnicas de aprendizagem de máquina em uma aplicação ainda pouco explorada como a classificação de documentos fiscais. O estudo de caso foi desenvolvido sobre um subconjunto de registros do banco de dados de Notas Fiscais Eletrônicas do Consumidor da Secretaria da Fazenda do RS, sobre os quais foram aplicadas rotinas de tonekização, remoção de *stop words* e *stemming*. Após o treinamento do classificador foram utilizadas, para análise dos resultados, métricas específicas e adequadas para as técnicas escolhidas, como acurácia, erro e metodologia de validação cruzada K-*Fold*, com $k = 10$ *folds*.

3.2 PROCEDIMENTOS METODOLÓGICOS

A idealização e treinamento de um classificador utilizando abordagem por aprendizado supervisionado seguiu os seguintes passos:

i. Pesquisa por técnicas eficientes de processamento de texto que possibilitassem lidar com grandes volumes de dados de modo a torná-los adequados para o treinamento de classificadores. Durante essa fase, foram estudados os trabalhos de Russel e Norvig (2003), no que diz respeito à aprendizagem de máquina e modelo de aprendizado supervisionado; o trabalho de Indurkhya e Damerau (2010), no que se refere a processamento de texto e métricas de extração de características; e o trabalho de Manning *et al.* (2008), no que se refere ao

algoritmo de classificação Naïve Bayes e metodologia de validação cruzada. Além disso, durante a pesquisa bibliográfica por trabalhos envolvendo classificação de texto para os códigos do sistema harmonizado (SH) feita em anais de conferências como *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, *IEEE International Conference on Data Mining* e *Conference on Artificial Intelligence*, e periódicos como *Journal of Machine Learning Research* e *Data mining and Knowledge Discovery*, tomou-se conhecimento do trabalho de Ding *et al.* (2015), que serviu como *baseline* para a comparação de desempenho com método desenvolvido ao longo desse trabalho (discutido na Seção 3.4).

ii. Estudo de técnicas de classificação de texto escaláveis e eficazes para a utilização de classificadores online. Essa pesquisa se deu sob a luz dos trabalhos de Manning et al. (2008), Leskovec *et al.* (2014) e Ding *et al*. (2015). No trabalho de Manning *et al*. (2008) foi justificado que o desempenho do Naïve Bayes é adequado para aplicações online e de tempo real por sua classificação ser realizada através do cálculo de uma fórmula fechada, em vez de métodos iterativos e aproximações de maior complexidade realizadas por outros algoritmos. Além disso, o espaço da estrutura de dados demandada pelo modelo é relativamente simples, consistindo nas probabilidades *a priori* e nas contagens de ocorrências de palavras em cada classe. No trabalho de Leskovec *et al.* (2014), intitulado *Mining of Massive Datasets*, os autores abordam especificamente a mineração de dados em bancos de dados muito grandes, e confiram que TF-IDF é uma métrica utilizada na atividade de classificação de textos quando se lida com grandes volumes de dados.

iii. Implementação de um classificador de texto que associe descrições textuais de itens de produtos aos seus respectivos códigos NCM adequados, através da utilização de aprendizagem de máquina em conjunto com o algoritmo de classificação Naïve Bayes.

iv. Definição dos experimentos a serem realizados e construção dos conjuntos de dados, baseados em dados reais de transações obtidas a partir do banco de dados da SEFAZ-RS, que serão utilizados para testes. A definição dos

experimentos foi baseada na estrutura de experimentos desenvolvidos no trabalho de Ding *et al.* (2015), onde optou-se por selecionar itens de NFC pertencentes aos códigos NCM dos Capítulos 22 e 90, de modo que fosse possível a comparação. A definição dos experimentos está detalhada mais adiante na Seção 3.4.

v. Teste do classificador sobre os conjuntos de dados definidos para os experimentos, através da apresentação das instâncias de teste rotuladas para posterior comparação do rótulo fornecido pelo classificador com o rótulo original da instância. Esse processo foi realizado em lote e utilizado em conjunto com a metodologia de validação cruzada com 10-*fold* para obtenção da acurácia, que é a medida escolhida para representar a proporção de instâncias corretamente classificadas em relação ao total de instâncias apresentadas.

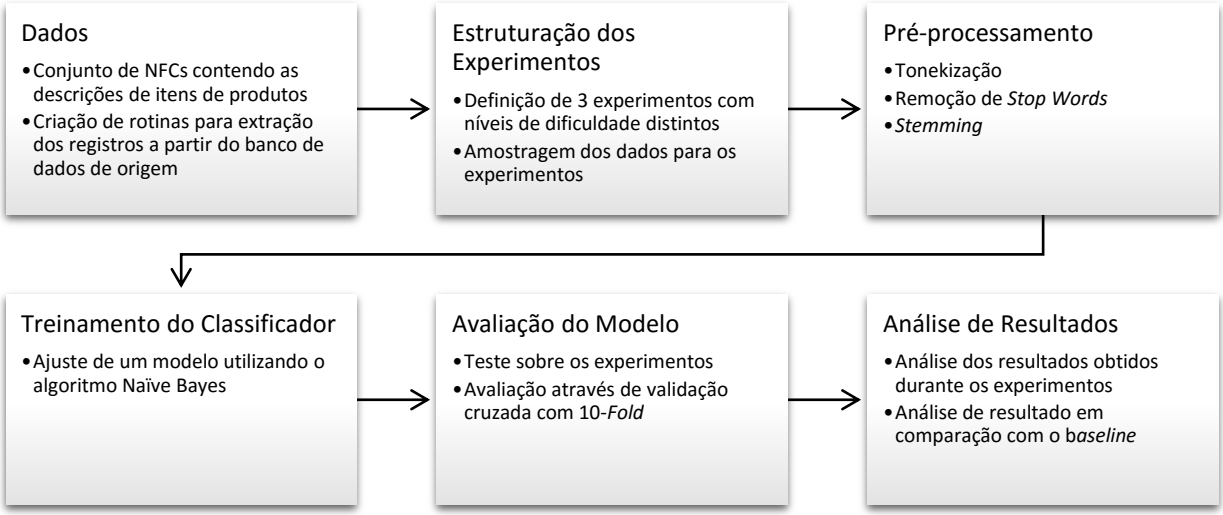
vi. Analise dos resultados do classificador implementado utilizando métricas adequadas para avaliação de modelos, a fim de validar o desempenho do classificador treinado. Para isso foram utilizados o cálculo do erro, o cálculo da acurácia e a aplicação da metodologia de validação cruzada para obtenção dessas medidas. A escolha pela acurácia se deu por essa refletir a quantidade de acertos em relação ao total de itens classificados, que é o objetivo da atividade desenvolvida nesse trabalho. Essa métrica, também presente na avaliação do classificador apresentado no trabalho de *Ding et al.* (2015), permitiu a comparação entre os modelos. Por outro lado, a opção pelo método de validação cruzada se deu pela necessidade de aproveitar melhor o conjunto de dados utilizado e com o objetivo de evitar qualquer possível viés que uma escolha viciada dos dados pudesse causar. Para isso, optou-se por utilizar K-*Fold* com $k = 10$, por esse ser uma escolha de partição não muito custosa, quando comparada ao *leave-one-out*, e utilizar 90% dos dados para treinamento e 10% dos dados para teste a cada rodada.

3.3 MÉTODO

O método utilizado para a construção de um classificador de descrições de itens de produtos em códigos NCM foi dividido em seis etapas, sendo essas: (1) extração dos

documentos NFC de sua base de origem; (2) definição dos experimentos; (3) pré-processamento dos documentos de entrada; (4) treinamento do classificador; (5) avaliação do modelo, que corresponde à classificação propriamente dita; e (6) análise dos resultados e comparação com o *baseline*. A Figura 3.1 ilustra as etapas e atividades que foram desenvolvidas durante o treinamento do classificador.

Figura 3.1 - Etapas para treinamento e avaliação do classificador.



Fonte: Autor.

O Algoritmo 1 descreve as etapas presentes no pré-processamento realizado sobre os documentos, e tem como objetivo extrair o vetor de características de cada um dos documentos de entrada. A rotina recebe um conjunto de documentos *D* como parâmetro, e inicia com tokenização dos documentos, quebrando o texto e extraindo vetor de *tokens* correspondente. O passo seguinte do algoritmo é remoção de *stop words*. A Tabela 3.1 apresenta a lista de *stop words* utilizadas nesse estudo. A sequência do pré-processamento se dá pela aplicação de *stemming* sobre os termos restantes nos documentos, conforme explicado na Subseção 2.3.3.

Algoritmo 1 - Algoritmo para Pré-processamento dos documentos

Entrada: Conjunto de Documentos (D)
Saída: Matriz (W) de pesos correspondente aos vetores de características extraídos de D

```
1.  ROTINA ExtrairVetorDeCaracteristicas(D)
```

```
 2.    Realizar tokenização nos documentos d_i ∈ D
 3.    Remover Stop Words de D
 4.    Realizar Stemming dos termos em D
 5.    Montar um histograma H com a frequência dos termos de D
 6.    Para cada d_i ∈ D:
 7.      Criar um vetor t contendo a frequência dos termos t_k ∈ d_i
 8.      Calcular TF-IDF dos termos t_k ∈ d_i em um vetor x usando t e H
 9.      Incluir x em W
10.    RETORNA W
```

Os passos seguintes do algoritmo de pré-processamento, representados nas linhas 5-9, referem-se à contabilização dos termos para calcular as métricas TF (Subseção 2.3.4) e IDF (Subseção 2.3.5), que servirão de base para o cálculo de TF-IDF (Subseção 2.3.6), cujos resultados produzirão os pesos dos vetores de características a serem utilizados para o treinamento do classificador.

A matriz de documentos pré-processados, composta pelos vetores de características, preenchidos com os pesos originados pelo TF-IDF, são passadas para o algoritmo de treinamento do classificador Naïve Bayes. O Algoritmo 2 descreve os passos necessários para o treinamento do classificador. O que se faz ao treinar um classificador Naïve Bayes é pré-computar as frequências de ocorrências de classes (linha 2), que serão utilizadas para cálculo das probabilidades *a priori*, e a matriz de ocorrência de *tokens* em classes (linha 3), que será utilizada no cálculo das probabilidades *a posteriori*, conforme descrito na Equação 6 da Subseção 2.2.2. O vetor H e a matriz F ficam disponíveis para o cálculo de probabilidades a ser realizado durante a etapa de classificação.

Algoritmo 2 – Algoritmo para treinamento do classificador Naïve Bayes.

Entrada: Conjunto de Descrições (X)
Conjunto de Rótulos (Y)
Saída: Histograma de Classes (H)
Matriz com a Frequência de Ocorrências dos Tokens x Classe (F)

```
 1.  ROTINA Treinar(X,Y)
 2.    Criar um histograma H com a frequência das classes c_i ∈ Y
 3.    Criar uma matriz F (t × c) de ocorrências de tokens x classes
 4.    RETORNA H, F
```

O Algoritmo 3 descreve os passos necessários para classificação de uma instância não rotulada x, utilizando as frequências pré-computadas armazenadas em H e F.

Nas linhas 2-3 são calculadas as probabilidades de x pertencer a cada uma das classes $c_i \in C$ possíveis. Após, em um segundo passo, é selecionada a classe com maior probabilidade $P(x, c_i)$ para que a classe correspondente seja retornada como rótulo pelo algoritmo.

Algoritmo 3 – Algoritmo para classificação de instâncias não rotuladas.

Entrada: Instância não rotulada (x)
Histograma de Classes (H)
Matriz com a Frequência de Ocorrências dos Tokens x Classe (F)
Saída: Rótulo de classe (y)

1. ROTINA Classificar(x)
2. Para cada $c_i \in H$:
3. Calcular a probabilidade de x pertencer à classe c_i $P(x, c_i)$, usando os valores computados em H e F
4. $y := \max_{c_i \in C} P(x, c_i)$
5. RETORNA y

Tabela 3.1 - Lista de *stop words* para língua portuguesa.

a	da	diz	feita	muito	pelos	quanto	tido
à	daquele	dizem	feitas	muitos	pequena	quantos	tinha
agora	daqueles	do	feito	na	pequenas	que	tinham
ainda	das	dos	feitos	não	pequeno	quem	toda
alguém	de	e	foi	nas	pequenos	são	todas
algum	dela	é	for	nem	per	se	todavia
alguma	delas	e'	foram	nenhum	perante	seja	todo
algumas	dele	ela	fosse	nessa	pode	sejam	todos
alguns	deles	elas	fossem	nessas	pôde	sem	tu
ampla	depois	ele	grande	nesta	podendo	sempre	tua
amplas	dessa	eles	grandes	nestas	poder	sendo	tuas
amplo	dessas	em	há	ninguém	poderia	será	tudo
amplos	desse	enquanto	isso	no	poderiam	serão	última
ante	desses	entre	isto	nos	podia	seu	últimas
antes	desta	era	já	nós	podiam	seus	último
ao	destas	essa	la	nossa	pois	si	últimos
aos	deste	essas	la	nossas	por	sido	um
após	deste	esse	lá	nosso	porém	só	uma
aquela	destes	esses	lhe	nossos	porque	sob	umas
aquelas	deve	esta	lhes	num	posso	sobre	uns
aquele	devem	está	lo	numa	pouca	sua	vendo
aqueles	devendo	estamos	mas	nunca	poucas	suas	ver
aquilo	dever	estão	me	o	pouco	talvez	vez
as	deverá	estas	mesma	os	poucos	também	vindo

até	deverão	estava	mesmas	ou	primeiro	tampouco	vir
através	deveria	estavam	mesmo	outra	primeiros	te	vos
cada	deveriam	estávamos	mesmos	outras	própria	tem	vós
coisa	devia	este	meu	outro	próprias	tendo	
coisas	deviam	estes	meus	outros	próprio	tenha	
com	disse	estou	minha	para	próprios	ter	
como	disso	eu	minhas	pela	quais	teu	
contra	disto	fazendo	muita	pelas	qual	teus	
contudo	dito	fazer	muitas	pelo	quando	ti	

Fonte: http://miningtext.blogspot.com.br/2008/11/listas-de-stopwords-stoplist-portugues.html.

3.4 DEFINIÇÃO DOS EXPERIMENTOS

Com o objetivo de testar o classificador desenvolvido nesse trabalho, foram desenvolvidos três experimentos que ocorreram sobre conjuntos de dados distintos compreendidos sobre transações que continham itens pertencentes a códigos de NCM dos capítulos 22 (Bebidas, líquidos alcoólicos e vinagres) e 90 (Instrumentos e aparelhos de óptica, de fotografia, de cinematografia, de medida, de controle ou de precisão; instrumentos e aparelhos médico-cirúrgicos; suas partes e acessórios). Segundo Ding *et al.* (2015), os itens pertencentes aos capítulos 22 e 90 do NCM são os que mais apresentam erros de classificação na prática e, portanto, são os que possuem maior probabilidade de erro. Essa possibilidade de erro pode se dar basicamente por dois fatores, sendo esses a quantidade de categorias pertencentes ao capítulo e a quantidade de tipos de produtos compreendidos. Assim, esses fatores contribuíram para que os capítulos 22 e 90 fossem escolhidos e utilizados para testar o desempenho do classificador.

Tendo como objetivo a avaliação do classificador com métricas e metodologia de validação aceitas pela comunidade científica e, com o objetivo de se ter um *baseline* para comparação da efetividade do método aplicado nesse estudo, optou-se por estruturar os experimentos tão próximos quanto o possível dos conjuntos de dados utilizados no trabalho de Ding *et al.* (2015). No trabalho mencionado, os autores construíram um classificador baseado em Background Nets para o Sistema Harmonizado (SH), e realizaram testes sobre descrições de itens que estavam no idioma inglês. Salvo as diferenças de idioma, o fato de o sistema NCM possuir hierarquia mais profunda do que o SH e, portanto, sua classificação ser mais específica, tentou-se tornar tão próximas quanto possível a quantidade de registros, a os códigos dos capítulos e métrica utilizada para avaliação.

3.4.1 SELEÇÃO DA AMOSTRA

Os registros de transações que deram origem aos conjuntos de dados utilizados nesse trabalho foram extraídos da base de Notas Fiscais do Consumidor da SEFAZ-RS e estavam armazenados em um servidor que utilizava tecnologia Microsoft SQL Server versão 2014. Com essa informação, pesquisou-se por recursos da linguagem T-SQL, extensão do padrão SQL ANSI, que pudessem ser úteis para a extração da amostra de dados.

Um dos cuidados que se tomou para a montagem da amostra, é que os dados fossem capturados de maneira puramente aleatória. Para isso, utilizou-se o recurso TABLESAMPLE, uma função nativa do T-SQL, que seleciona registros de maneira aleatória. Essa função recebe como parâmetro a quantidade de registros a serem amostrados, podendo essa quantidade ser informada tanto em valor absoluto, como em percentual.

O Algoritmo 4 apresenta a consulta utilizada para a seleção das transações contendo descrições de produtos correspondentes aos Capítulos 22 e 90 do NCM, que foram utilizadas para calibragem e validação do modelo.

Algoritmo 4 - Consulta para seleção dos itens pertencentes aos Capítulos 22 e 90 do NCM.

```
1.  SELECT COD_NCM, TEX_DESC_PSERV, QTD_TRIB, UNID_TRIB
2.    FROM ADT_NFE_DF_ITEM A TABLESAMPLE (2330000 ROWS)
3.    WHERE LEFT(COD_NCM, 2) = '22'
4.
5.  SELECT COD_NCM, TEX_DESC_PSERV, QTD_TRIB, UNID_TRIB
6.    FROM ADT_NFE_DF_ITEM A TABLESAMPLE (3400000 ROWS)
7.    WHERE LEFT(COD_NCM, 2) = '90'
```

Após a seleção da amostra foi realizado uma análise com o objetivo de eliminar documentos que apresentassem descrições de itens duplicadas, assim como a contagem das classes existentes em cada conjunto de dados. A Tabela 3.2 apresenta o total de registros presente em cada uma das amostras de NFC extraídas antes e depois da remoção de duplicatas, assim como a quantidade de categorias existentes em cada um. Nesse ponto, duplicatas referem-se a itens de produtos que apresentavam descrições repetidas.

Tabela 3.2 - Número de registros e categorias antes e depois do pré-processamento.

Dataset, Registros, Categorias	Antes	Depois
A - Dataset Capítulo 22 (1)		
Registros	210.836	20.259
Categorias	29	29
B - Dataset Capítulo 22 (2)		
Registros	207.310	64.269
Categorias	59	59
C - Dataset Capítulo 90		
Registros	412.969	262.244
Categorias	819	819

3.4.2 AVALIAÇÃO DO MODELO

Para a realização da avaliação do modelo utilizou-se a acurácia média calculada sobre os testes parciais do método de validação cruzada K-fold com $k = 10$ *folds*, conforme tratado na Subseção 2.4.1. Apesar de existirem outras métricas de avaliação de desempenho utilizadas nas áreas de classificação de texto e recuperação de informações, tais como precisão, revocação e a média harmônica entre essas, denominada medida-F1, a métrica mais relevante para avaliar os resultados de modelos de classificação de NCM mostrou-se a acurácia (DING *et al.*, 2015). A acurácia apresenta a proporção entre as instâncias que o modelo classificou corretamente e o total de instâncias classificadas pelo modelo.

A adoção do método de avaliação cruzada com 10-*folds*, diferente do método adotado no estudo utilizado como *baseline*, que realiza a divisão dos dados em conjuntos de 60% dos dados para treinamento e 40% para teste, auxilia em eliminar algum viés realizado por uma eventual escolha tendenciosa dos dados. Apesar de ser mais custoso computacionalmente, o método de avaliação cruzada com 10-*folds* auxilia na redução da variância (KOHAVI, 1995), além de permitir uma avaliação mais precisa da acurácia.

3.5 IMPLEMENTAÇÃO

O classificador Naïve Bayes desenvolvido ao longo desse estudo foi implementado

utilizando a linguagem de programação Python 2.7[1]. Para as tarefas relacionadas ao processamento de texto, utilizou-se a biblioteca NLTK (BIRD *et al.*, 2009). A biblioteca NLTK é uma suíte para processamento de linguagem natural que possui mais de 50 corpora de exemplo e recursos para classificação de texto, *tokenização*, *stemming*, rotulação, *parsing* e associação semântica. O pacote NLTK pode ser utilizado também para tradução de máquina, rotulação semântica, análise de sentimento, modelagem de tópicos e possui funções para download e processamento de mensagens do Twitter.

O treinamento do classificador bayesiano, assim como a metodologia de avaliação, baseada em validação cruzada, foram implementados com o suporte das rotinas presentes na biblioteca Scikit-learn (PEDREGOSA *et al*., 2011). Scikit-learn é uma biblioteca escrita em Python e que recentemente tornou-se difundida entre os entusiastas e cientistas de dados sendo até mesmo utilizada em sistemas em produção. Essa biblioteca é utilizada para desenvolvimento de classificadores (como detecção de spam, reconhecimento de imagem, etc.), análise de regressão (como resposta a medicamentos, preço de ações, etc.), agrupamento (como segmentação de clientes, agrupamento de notícias por assunto, etc.), redução de dimensionalidade (como visualização, aumento de eficiência, etc.), seleção de modelos (como comparação, validação, seleção de parâmetros, etc.) e pré-processamento de dados de entrada (extração de características, normalização, etc.).

Dos recursos mencionados, o pré-processamento da entrada, *tokenização*, remoção de *stop words*, treinamento do modelo de classificação, estudo dos parâmetros e avaliação de desempenho utilizando validação cruzada, foram realizadas com o auxílio da biblioteca Scikit-learn. Entre os recursos oferecidos pela biblioteca NLTK, apenas o recurso de *stemming* foi utilizado.

3.6 CONSIDERAÇÕES

Nesse capítulo foi apresentada a metodologia adotada para o desenvolvimento de um classificador de NCM com base em descrições de itens de produtos. Foram descritas as etapas envolvidas no pré-processamento da entrada, treinamento do modelo, classificação e exibidos os algoritmos em pseudo-código que implementam os conceitos introduzidos na fundamentação teórica (Capítulo 2).

[1] http://www.python.org/

Também foram descritos os conjuntos de dados utilizados nos experimentos, como se deu sua idealização e um breve histórico do caso a ser utilizado como *baseline* durante a etapa de análise dos resultados. Além disso, descreveu-se o procedimento para seleção amostral dos registros de transações através de recursos específicos da linguagem SQL e especificadas as métricas e metodologia a serem utilizadas durante a validação do modelo, que utilizou a acurácia geral junto da abordagem de validação cruzada com 10-*folds*.

Por fim, foram descritas as tecnologias empregadas na implementação do modelo, todas baseadas em software livre. A linguagem Python, em conjunto com os pacotes NLTK e Scikit-learn foram às escolhas adotadas durante o desenvolvimento, teste e validação do modelo.

4. APRESENTAÇÃO DOS RESULTADOS E DISCUSSÕES

4.1 EXPERIMENTO SOBRE O CONJUNTO DE DADOS A

O primeiro experimento foi realizado sobre o conjunto de dados baseado em 210.836 transações ocorridas no ano de 2016, pertencentes a 29 NCMs do Capítulo 22 (Bebidas, líquidos alcoólicos e vinagres), selecionadas de modo aleatório, de uma grande rede varejista do Rio Grande do Sul. Por se tratar de registros de uma mesma rede de varejo, pressupõe-se que as descrições de itens sigam certa uniformidade, o que motivou o julgamento da classificação desse conjunto como fácil. Sobre esse conjunto de treinamento foi realizado o processamento de texto a fim de se obter os vetores de características para cada instância e o pré-processamento dessas através do cálculo do TF-IDF. O conjunto de instâncias foi apresentado para o treinamento do classificador Naïve Bayes Multinomial.

Uma vez treinado o modelo do classificador, prosseguiu-se então com a avaliação de desempenho do modelo, em relação a sua capacidade em fazer classificações corretas, utilizando-se a técnica de validação cruzada com 10-*folds*. A Tabela 4.1 exibe o resultado do modelo para o conjunto de dados em questão. Nessa tabela é possível observar que o resultado da acurácia do classificador para o conjunto com duplicatas é igual a 99,55% de acerto, enquanto que para o conjunto sem duplicatas, a acurácia ficou em 98,68%. Esse resultado superior para o conjunto de dados com duplicatas sugere que a maior parte das duplicatas referem-se a instâncias para as quais o classificador correlacionou corretamente, pois contribuíram positivamente para a métrica.

Tabela 4.1 - Acurácia média do classificador sobre o conjunto de dados A para os casos com e sem duplicatas.

Conjunto de Dados	Configuração	Média
A - Dataset Capítulo 22 (1)	com Duplicatas	0,9955
	sem Duplicatas	0,9868

4.2 EXPERIMENTO SOBRE O CONJUNTO DE DADOS B

O segundo experimento foi realizado sobre um conjunto de dados baseado em 207.310

registros de transações ocorridas no ano de 2016, distribuídas em 29 códigos distintos de NCMs e pertencentes, assim como o conjunto de dados anterior, ao Capítulo 22. No entanto, diferente do conjunto de dados do experimento anterior, as transações selecionadas nesse caso foram de estabelecimentos comerciais aleatoriamente selecionados. Com isso, espera-se que as descrições de itens, apesar de pertencerem todas ao mesmo Capítulo, possuam maiores divergências e variações em suas descrições, motivando a classificação desse conjunto de dados como dificuldade média.

Quando treinou-se o classificador e prosseguiu-se com a avaliação de performance do modelo, obteve-se os resultados apresentados na Tabela 4.2. Confirmando a suspeita de tratar-se de um conjunto de dados mais difícil do que o utilizado no experimento anterior, observou-se uma queda na performance do classificador. Mais uma vez o conjunto de dados com duplicatas apresentou uma performance melhor, atingindo uma acurácia média de 95,60%. Para o conjunto de dados sem duplicatas, o desempenho do classificador atingiu 90,30%.

Tabela 4.2 - Acurácia média do classificador sobre o conjunto de dados B para os casos com e sem duplicatas.

Conjunto de Dados	Configuração	Média
B - Dataset Capítulo 22 (2)	com Duplicatas	0,9560
	sem Duplicatas	0,9030

4.3 EXPERIMENTO SOBRE O CONJUNTO DE DADOS C

O terceiro e último experimento foi realizado sobre um conjunto de dados baseado em 412.969 registros de transações ocorridas no ano de 2016 e distribuídas em 819 códigos distintos de NCMs, pertencentes ao Capítulo 90. As transações selecionadas para esse experimento pertenciam a estabelecimentos comerciais do Rio Grande do Sul, aleatoriamente selecionados. Além disso, existe um aumento na complexidade desse conjunto de dados, devido ao fato da quantidade elevada de classes (códigos NCM). Com isso, espera-se que as descrições de itens, apesar de ainda pertencerem todas ao um mesmo capítulo, sejam mais difíceis de serem classificadas, devido à especificidade da hierarquia de classes, fazendo com que esse conjunto de dados fosse classificado como difícil.

A Tabela 4.3 exibe os resultados do desempenho do classificador para o terceiro

experimento. Conforme o esperado, por se tratar de um conjunto de dados mais difícil, observou-se uma queda nos valores de acurácia, tanto para o caso com duplicatas (88,22%), quanto para o caso sem duplicatas (83,38%), permanecendo o caso com duplicatas apresentando performance superior.

Tabela 4.3 - Acurácia média do classificador sobre o conjunto de dados C para os casos com e sem duplicatas.

Conjunto de Dados	Configuração	Média
C - Dataset Capítulo 90	com Duplicatas	0,8822
	sem Duplicatas	0,8338

4.4 COMPARAÇÃO ENTRE RESULTADOS OBTIDOS COM *BASELINE*

Ao confrontar os resultados obtidos durante a avaliação do classificador com os resultados do estudo definido como *baseline* na Seção 3.4, observou-se que os resultados ficaram ligeiramente abaixo ao comparar o conjunto de dados B, sem duplicatas, da Tabela 4.4 (90,30%), com o conjunto de dados X, da Tabela 4.5 (98,56%), que correspondem aos experimentos sobre o Capítulo 22. No entanto, quando se comparou os experimentos relacionados ao Capítulo 90, que corresponde ao conjunto de dados difícil, notou-se um desempenho equivalente. Os resultados do classificador desenvolvido nesse estudo estão exibidos no conjunto de dados C, sem duplicatas, da Tabela 4.4 (83,38%), enquanto que os dados do *baseline* estão dispostos no conjunto de dados Y da Tabela 4.5 (83,30%).

Tabela 4.4 - Acurácia média do classificador para os conjuntos de dados A, B e C para os casos com e sem duplicatas.

Conjunto de Dados	Configuração	Média
A - Dataset Capítulo 22 (1)	com Duplicatas	0,9955
	sem Duplicatas	0,9868
B - Dataset Capítulo 22 (2)	com Duplicatas	0,9560
	sem Duplicatas	0,9030
C - Dataset Capítulo 90	com Duplicatas	0,8822

	sem Duplicatas	0,8338

Tabela 4.5 - Acurácia Média dos Capítulos 22 e 90 do SH para o classificador desenvolvido por Ding *et al.* (2015).

Conjunto de Dados	Configuração	Média
X – Dataset Capítulo 22 do SH	sem Duplicatas	0,9856
Y – Dataset Capítulo 90 do SH	sem Duplicatas	0,8330

Entre os fatores que podem justificar o desempenho inferior para o primeiro caso (conjunto de dados B x conjunto de dados X), está o fato de se estar comparando dados do sistema NCM com dados do sistema harmonizado. Enquanto o sistema harmonizado possui 6 dígitos, o sistema NCM possui 8 dígitos, sendo uma hierarquia mais profunda e, portanto, mais específica, fazendo com que a mesma tenha mais classes e, portanto, seja mais difícil de classificar. Isso pode ser constatado, pois no conjunto de dados desse experimento, o Capítulo 22 possui 59 classes, versus 52 classes existentes no *baseline*, enquanto que para o Capítulo 90, o conjunto de dados desse experimento possui 819 classes, versus 204 classes do conjunto de dados utilizado como *baseline*.

Outro fator que pode explicar a diferença de desempenho entre os classificadores é o fato do experimento desenvolvido nesse estudo estar utilizando descrições de produtos escritas no idioma Português, enquanto que os conjuntos de dados do *baseline* possuíam descrições em inglês. Esse fator é muito importante e não pode ser ignorado, uma vez que a língua portuguesa é mais complexa e mais rica em variações linguísticas sendo, certamente, mais difícil de classificar.

Além disso, outro fator a ser questionado é se está correta a remoção das descrições de produtos duplicadas, uma vez que se está utilizando um classificador com base estatística, que armazena as probabilidades *a priori* de cada classe e encontrará essa mesma distribuição de valores quando tornar a olhar os registros de transações do mundo real. Como o método de seleção dos dados é aleatório, o fato de existir uma distribuição desigual de itens entre as classes está evidenciando o fato de que aquela classe é mais provável de ocorrer, que é uma característica do conjunto de dados original que está sendo representada na amostra.

CONCLUSÃO

O problema abordado ao longo deste trabalho foi a classificação automática de descrições textuais de itens de produtos em seus respectivos códigos NCM, motivado pelo objetivo de identificar as divergências existentes entre os códigos informados pelos estabelecimentos comerciais e os códigos reais aos quais os produtos pertencem. Esse estudo envolveu pesquisa relacionada a técnicas de processamento de texto, treinamento de classificadores utilizando aprendizado supervisionado e métodos de avaliação.

O classificador desenvolvido ao longo desse trabalho utilizou o algoritmo Naïve Bayes para a construção de seu modelo de classificação. O método proposto foi testado em três conjuntos de dados experimentais: o primeiro deles, considerado fácil, composto por transações pertencentes a NCMs do Capítulo 22 de uma grande rede varejista; o segundo conjunto de dados, considerado de dificuldade intermediária, composto por descrições de itens pertencentes a NCMs do Capítulo 22, pertencentes a transações de estabelecimentos comerciais selecionados aleatoriamente; e, por fim, o terceiro conjunto, considerado difícil por conter descrições de itens de NCMs do Capítulo 90, pertencentes a estabelecimentos comerciais amostrados aleatoriamente.

Os conjuntos de dados foram pré-processados com o objetivo de extrair os vetores de características utilizados como entrada para o algoritmo de treinamento do classificador Naïve Bayes. Esse pré-processamento envolveu *tokenização*, remoção de *stop words*, *stemming* e cálculo dos pesos dos vetores de características utilizando a métrica TF-IDF. Em seguida, foi realizado o treinamento do classificador, teste e avaliação do desempenho do mesmo utilizando a metodologia de validação cruzada.

Os resultados foram analisados e evidenciaram as características de dificuldade dos conjuntos de dados selecionados como experimento. Quando os resultados dos experimentos foram comparados com o caso utilizado como *baseline*, discutido no trabalho de Ding *et al.* (2015), observou-se que para os itens do Capítulo 22, o classificador desenvolvido nesse estudo mostrou desempenho inferior ao apresentado pelo *baseline*. No entanto, ao se considerar conjunto de dados difícil, composto por transações de 819 classes (códigos de NCMs distintos), observaram-se resultados equivalentes. Do ponto de vista prático, considera-se que os resultados obtidos são significativos, uma vez que o sistema NCM, por possuir uma hierarquia mais profunda e, portanto, específica, é mais difícil de classificar do que o sistema harmonizado (SH), que possui uma hierarquia mais rasa. Outro fator a ser

considerado, e que poderia contribuir para essa diferença de desempenho, é o fato desse estudo estar utilizando descrições de produtos baseadas na língua portuguesa, enquanto que os experimentos do *baseline* foram baseados na língua inglesa, que é mais simples e regular sendo, portanto, mais fácil de classificar.

Em relação aos objetivos iniciais que motivaram o desenvolvimento desse trabalho, acredita-se que tenham sido satisfatoriamente atingidos, uma vez que conseguiu-se estudar os conceitos e métodos relacionados à aprendizagem de máquina e processamento de texto e aplicá-los com sucesso no desenvolvimento de um classificador de NCM. Além disso, o classificador foi validado e comparado com outro trabalho da literatura, reforçando os resultados aqui descritos. Dessa forma, o classificador especificado ao longo desse trabalho pode ser utilizado na classificação de descrições de itens de produtos em códigos NCM, além de contribuir para a identificação de divergências existentes em notas fiscais do consumidor emitidas pelos estabelecimentos comerciais, auxiliando assim na atividade de auditoria eletrônica.

SUGESTÕES DE TRABALHOS FUTUROS

Apesar dos modelos desenvolvidos ao longo desse trabalho apresentarem desempenho satisfatório sobre Capítulos do NCM relativamente difíceis de classificar, o desenvolvimento da solução completa depende do treinamento de classificadores para os demais tipos de documentos. Para que isso seja possível, também há a necessidade de se dispor de conjuntos de dados rotulados ainda maiores, de modo que possam ser utilizados para o treinamento dos demais modelos.

Algoritmos baseados em aprendizagem de máquina normalmente utilizam conjuntos de dados selecionados aleatoriamente para o treinamento dos modelos. Em muitos desses casos, existe a possibilidade de utilizar aprendizado ativo, que permite ao algoritmo sinalizar ao especialista as instâncias que, se rotuladas, aumentam o desempenho do classificador. Esse tipo de técnica permite que após o treinamento inicial do modelo, esse seja aperfeiçoado ao longo do seu uso, através do fornecimento de instâncias rotuladas que garantam o aumento de sua acurácia. Mais detalhes sobre a utilização de aprendizado ativo em classificação de texto estão disponíveis no trabalho de Tong e Koller (2001).

Além disso, apesar da construção de um classificador de NCM auxiliar na tarefa de categorizar produtos com base em suas descrições, trata-se de apenas uma atividade inserida

em um contexto maior, que é a aplicação desse classificador em larga escala sobre bancos de dados massivos. Assim, outra sugestão de trabalho futuro é o desenvolvimento de tal arquitetura, que além de permitir a classificação de grandes quantidades de documentos em lote, também auxiliaria na atividade de auditoria eletrônica, através da identificação de divergências entre os códigos NCM informados pelos estabelecimentos comerciais e os códigos inferidos pelo classificador treinado.

REFERÊNCIAS

BIRD, Steven; Loper, Edward; Klein, Ewan (2009), Natural Language Processing with Python. O'Reilly Media Inc.

DING, Liya; FAN, ZhenZhen; CHEN, DongLiang. Auto-Categorization of HS Code Using Background Net Approach. Procedia Computer Science, v. 60, p. 1462-1471, 2015.

INDURKHYA, Nitin; DAMERAU, Fred J. (Ed.). Handbook of natural language processing. CRC Press, 2010.

JAIN, Raj. The art of computer systems performance analysis. John Wiley & Sons, 2008.

KORDE, Vandana, and C. Namrata Mahender. Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications (IJAIA) 3.2 (2012): 85-99.

KO, Youngjoong, and Jungyun Seo. Automatic text categorization by unsupervised learning. Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2000.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint Conference on artificial intelligence. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

KOHONEN, Teuvo. The self-organizing map. Neurocomputing 21.1 (1998): 1-6.

LESKOVEC, Jure; RAJARAMAN, Anand; ULLMAN, Jeffrey David. Mining of massive datasets. Cambridge University Press, 2014.

LUHN, Hans Peter. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, v. 1, n. 4, p. 309-317, 1957.

MANNING, Christopher D. et al. Introduction to information retrieval. Cambridge: Cambridge university press, 2008.

MCCALLUM, Andrew. Multi-label text classification with a mixture model trained by EM. AAAI'99 workshop on text learning. 1999.

MINISTÉRIO DA INDÚSTRIA, COMÉRCIO EXTERIOR E SERVIÇOS. TEC em Excel Completa, 2016. Disponível em: <http://www.mdic.gov.br/comercio-exterior/estatisticas-de-comercio-exterior-9/arquivos-atuais>. Acesso em: 20 de out. 2016.

MITCHELL, Tom M. Machine learning. 1997. Burr Ridge, IL: McGraw Hill, v. 45, p. 37, 1997.

ORAIR, Gustavo Henrique. "Classificação multi-rótulo hierárquica de documentos textuais." (2009).

ORENGO, Viviane Moreira; HUYCK, Christian R. RSLP Stemmer (Removedor de Sufixos da Língua Portuguesa). Disponível em http://www.inf.ufrgs.br/~viviane/rslp/. Acesso em: 17 de set. 2016

ORENGO, Viviane Moreira; HUYCK, Christian R. A Stemming Algorithmm for the Portuguese Language. In: spire. 2001. p. 186-193.

PEDREGOSA et al. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.

RECEITA FEDERAL. Sistema harmonizado de designação e de codificação de mercadorias, 2015. Disponível em: <http://idg.receita.fazenda.gov.br/acesso-rapido/legislacao/legislacao-por-assunto/sistema-harmonizado>. Acesso em: 16 de dez. 2016.

RUSSELL, Stuart Jonathan; NORVIG, Peter. Artificial intelligence: a modern approach. Upper Saddle River: Prentice hall, 2003.

SAJNANI, Hitesh, et al. "Multi-Label Classification of Short Text: A Study on Wikipedia Barnstars." Analyzing Microtext. 2011.

SPARCK JONES, Karen. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, v. 28, n. 1, p. 11-21, 1972.

TONG, Simon; KOLLER, Daphne. Support vector machine active learning with applications to text classification. Journal of machine learning research, v. 2, n. Nov, p. 45-66, 2001.

TRINKLE, Patrick. An introduction to unsupervised document classification. (2009).

TRIOLA, Mario F. Bayes' Theorem, 2008. Disponível em: < http://faculty.washington.edu/tamre/BayesTheorem.pdf>. Acesso em: 11 de out. 2016.

YANG, Bishan, et al. "Effective multi-label active learning for text classification." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

ZAKI, Mohammed J.; MEIRA JR, Wagner; MEIRA, Wagner. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.

ANEXO A – Códigos NCM do Capítulo 22

Tabela A.1 - Tabela de NCM do Capítulo 20 (Bebidas, líquidos alcoólicos e vinagres).

NCM	DESCRIÇÃO	TEC (%)
22.01	**Águas, incluindo as águas minerais, naturais ou artificiais, e as águas gaseificadas, não adicionadas de açúcar ou de outros edulcorantes nem aromatizadas; gelo e neve.**	
2201.10.00	- Águas minerais e águas gaseificadas	20
2201.90.00	- Outros	20
22.02	**Águas, incluindo as águas minerais e as águas gaseificadas, adicionadas de açúcar ou de outros edulcorantes ou aromatizadas e outras bebidas não alcoólicas, exceto sucos (sumos) de fruta ou de produtos hortícolas, da posição 20.09.**	
2202.10.00	- Águas, incluindo as águas minerais e as águas gaseificadas, adicionadas de açúcar ou de outros edulcorantes ou aromatizadas	20
2202.9	- Outras:	
2202.91.00	-- Cerveja sem álcool	20
2202.99.00	-- Outras	20
2203.00.00	**Cervejas de malte.**	20
22.04	**Vinhos de uvas frescas, incluindo os vinhos enriquecidos com álcool; mostos de uvas, excluindo os da posição 20.09.**	
2204.10	- Vinhos espumantes e vinhos espumosos	
2204.10.10	Tipo champanha (*champagne*)	20
2204.10.90	Outros	20
2204.2	- Outros vinhos; mostos de uvas cuja fermentação tenha sido impedida ou interrompida por adição de álcool:	
2204.21.00	-- Em recipientes de capacidade não superior a 2 l	20#
2204.22	-- Em recipientes de capacidade superior a 2 l, mas não superior a 10 l	
2204.22.1	Vinhos	
2204.22.11	Em recipientes de capacidade não superior a 5 l	20
2204.22.19	Outros	20
2204.22.20	Mostos	20
2204.29	-- Outros	
2204.29.10	Vinhos	20
2204.29.20	Mostos	20
2204.30.00	- Outros mostos de uvas	20
22.05	**Vermutes e outros vinhos de uvas frescas aromatizados por plantas ou substâncias aromáticas.**	
2205.10.00	- Em recipientes de capacidade não superior a 2 l	20
2205.90.00	- Outros	20

2206.00	**Outras bebidas fermentadas (por exemplo, sidra, perada, hidromel, saquê); misturas de bebidas fermentadas e misturas de bebidas fermentadas com bebidas não alcoólicas, não especificadas nem compreendidas noutras posições.**	
2206.00.10	Sidra	20
2206.00.90	Outras	20
22.07	**Álcool etílico não desnaturado, com um teor alcoólico, em volume, igual ou superior a 80 % vol; álcool etílico e aguardentes, desnaturados, com qualquer teor alcoólico.**	
2207.10	- Álcool etílico não desnaturado, com um teor alcoólico, em volume, igual ou superior a 80 % vol	
2207.10.10	Com um teor de água inferior ou igual a 1 % vol	20#
2207.10.90	Outros	20
2207.20	- Álcool etílico e aguardentes, desnaturados, com qualquer teor alcoólico	
2207.20.1	Álcool etílico	
2207.20.11	Com um teor de água inferior ou igual a 1 % vol	20#
2207.20.19	Outros	20
2207.20.20	Aguardente	20
22.08	**Álcool etílico não desnaturado, com um teor alcoólico, em volume, inferior a 80 % vol; aguardentes, licores e outras bebidas espirituosas.**	
2208.20.00	- Aguardentes de vinho ou de bagaço de uvas	20
2208.30	- Uísques	
2208.30.10	Com um teor alcoólico, em volume, superior a 50 % vol, em recipientes de capacidade igual ou superior a 50 l	12
2208.30.20	Em embalagens de capacidade inferior ou igual a 2 l	20
2208.30.90	Outros	20
2208.40.00	- Rum e outras aguardentes provenientes da destilação, após fermentação, de produtos da cana-de-açúcar	20
2208.50.00	- Gim e genebra	20
2208.60.00	- Vodca	20
2208.70.00	- Licores	20
2208.90.00	- Outros	20
2209.00.00	**Vinagres e seus sucedâneos obtidos a partir do ácido acético, para usos alimentares.**	20

ANEXO B – Códigos NCM do Capítulo 90

Tabela B.1 - Tabela de NCM do Capítulo 90 (Instrumentos e aparelhos de óptica, de fotografia, de cinematografia, de medida, de controle ou de precisão; instrumentos e aparelhos médico-cirúrgicos; suas partes e acessórios).

NCM	DESCRIÇÃO	TEC (%)
90.01	**Fibras ópticas e feixes de fibras ópticas; cabos de fibras ópticas, exceto os da posição 85.44; matérias polarizantes em folhas ou em placas; lentes (incluindo as de contato), prismas, espelhos e outros elementos de óptica, de qualquer matéria, não montados, exceto os de vidro não trabalhado opticamente.**	
9001.10	- Fibras ópticas, feixes e cabos de fibras ópticas	
9001.10.1	Fibras ópticas	
9001.10.11	Com diâmetro de núcleo inferior a 11 micrômetros (mícrons)	12BIT
9001.10.19	Outras	12BIT
9001.10.20	Feixes e cabos de fibras ópticas	14BIT
9001.20.00	- Matérias polarizantes, em folhas ou em placas	18
9001.30.00	- Lentes de contato	18
9001.40.00	- Lentes de vidro, para óculos	18
9001.50.00	- Lentes de outras matérias, para óculos	18
9001.90	- Outros	
9001.90.10	Lentes	18
9001.90.90	Outros	18
90.02	**Lentes, prismas, espelhos e outros elementos de óptica, de qualquer matéria, montados, para instrumentos ou aparelhos, exceto os de vidro não trabalhado opticamente.**	
9002.1	- Objetivas:	
9002.11	-- Para câmeras, para projetores ou para aparelhos fotográficos ou cinematográficos, de ampliação ou de redução	
9002.11.10	Para câmeras fotográficas ou cinematográficas ou para projetores	16
9002.11.20	De aproximação (*zoom*) para câmeras de televisão, de 20 ou mais aumentos	0BK
9002.11.90	Outras	16
9002.19.00	-- Outras	16
9002.20	- Filtros	
9002.20.10	Polarizantes	16
9002.20.90	Outros	16
9002.90.00	- Outros	16
90.03	**Armações para óculos e artigos semelhantes, e suas partes.**	
9003.1	- Armações:	
9003.11.00	-- De plásticos	18
9003.19	-- De outras matérias	
9003.19.10	De metais comuns, mesmo folheados ou chapeados de metais preciosos (plaquê)	18
9003.19.90	Outras	18
9003.90	- Partes	
9003.90.10	Charneiras	16

9003.90.90	Outras	16
90.04	**Óculos para correção, proteção ou outros fins, e artigos semelhantes.**	
9004.10.00	- Óculos de sol	20
9004.90	- Outros	
9004.90.10	Óculos para correção	18
9004.90.20	Óculos de segurança	18
9004.90.90	Outros	18
90.05	**Binóculos, lunetas, incluindo as astronômicas, telescópios ópticos, e suas armações; outros instrumentos de astronomia e suas armações, exceto os aparelhos de radioastronomia.**	
9005.10.00	- Binóculos	18
9005.80.00	- Outros instrumentos	14BK
9005.90	- Partes e acessórios (incluindo as armações)	
9005.90.10	De binóculos	16
9005.90.90	Outros	14BK
90.06	**Câmeras fotográficas; aparelhos e dispositivos, incluindo as lâmpadas e tubos, de luz-relâmpago (flash) para fotografia, exceto as lâmpadas e tubos de descarga da posição 85.39.**	
9006.10	- Câmeras fotográficas dos tipos utilizados para preparação de clichês ou cilindros de impressão	
9006.10.10	Fotocompositoras a laser para preparação de clichês	0BK
9006.10.90	Outras	14BK
9006.30.00	- Câmeras fotográficas especialmente concebidas para fotografia submarina ou aérea, para exame médico de órgãos internos ou para laboratórios de medicina legal ou de investigação judicial	14BK
9006.40.00	- Câmeras fotográficas para filmes de revelação e copiagem instantâneas	2
9006.5	- Outras câmeras fotográficas:	
9006.51.00	-- Com visor de reflexão através da objetiva (*reflex*), para filmes em rolos de largura não superior a 35 mm	18
9006.52.00	-- Outras, para filmes em rolos de largura inferior a 35 mm	18
9006.53	-- Outras, para filmes em rolos de 35 mm de largura	
9006.53.10	De foco fixo	18
9006.53.20	De foco ajustável	18
9006.59	-- Outras	
9006.59.10	De foco fixo	18
9006.59.2	De foco ajustável	
9006.59.21	Para obtenção de negativos de 45 mm x 60 mm ou de dimensões superiores	10
9006.59.29	Outras	18
9006.6	- Aparelhos e dispositivos, incluindo as lâmpadas e tubos, de luz-relâmpago (flash) para fotografia:	
9006.61.00	-- Aparelhos de tubo de descarga para produção de luz-relâmpago (denominados "flashes eletrônicos")	18
9006.69.00	-- Outros	18
9006.9	- Partes e acessórios:	
9006.91	-- De câmeras fotográficas	
9006.91.10	Corpos	16
9006.91.90	Outros	16

9006.99.00	-- Outros	16
90.07	**Câmeras e projetores, cinematográficos, mesmo com aparelhos de gravação ou de reprodução de som incorporados.**	
9007.10.00	- Câmeras	14BK
9007.20	- Projetores	
9007.20.20	Para filmes de largura superior ou igual a 35 mm mas inferior ou igual a 70 mm	0BK
9007.20.90	Outros	14BK
9007.9	- Partes e acessórios:	
9007.91.00	-- De câmeras	14BK
9007.92.00	-- De projetores	14BK
90.08	**Aparelhos de projeção fixa; aparelhos fotográficos de ampliação ou de redução.**	
9008.50.00	- Projetores e aparelhos de ampliação ou de redução	18
9008.90.00	- Partes e acessórios	16
90.10	**Aparelhos e material dos tipos usados nos laboratórios fotográficos ou cinematográficos, não especificados nem compreendidos noutras posições do presente Capítulo; negatoscópios; telas para projeção.**	
9010.10	- Aparelhos e material para revelação automática de filmes fotográficos, de filmes cinematográficos ou de papel fotográfico, em rolos, ou para copiagem automática de filmes revelados em rolos de papel fotográfico	
9010.10.10	Cubas e cubetas, de operação automática e programáveis	0BK
9010.10.20	Ampliadoras-copiadoras automáticas para papel fotográfico, com capacidade superior a 1.000 cópias por hora	0BK
9010.10.90	Outros	14BK
9010.50	- Outros aparelhos e material para laboratórios fotográficos ou cinematográficos; negatoscópios	
9010.50.10	Processadores fotográficos para o tratamento eletrônico de imagens, mesmo com saída digital	0BK
9010.50.20	Aparelhos para revelação automática de chapas de fotopolímeros com suporte metálico	0BK
9010.50.90	Outros	18
9010.60.00	- Telas para projeção	18
9010.90	- Partes e acessórios	
9010.90.10	De aparelhos ou material da subposição 9010.10 ou do item 9010.50.10	14BK
9010.90.90	Outros	16
90.11	**Microscópios ópticos, incluindo os microscópios para fotomicrografia, cinefotomicrografia ou microprojeção.**	
9011.10.00	- Microscópios estereoscópicos	14BK
9011.20	- Outros microscópios, para fotomicrografia, cinefotomicrografia ou microprojeção	
9011.20.10	Para fotomicrografia	0BK
9011.20.20	Para cinefotomicrografia	0BK
9011.20.30	Para microprojeção	0BK
9011.80	- Outros microscópios	
9011.80.10	Binoculares de platina móvel	0BK
9011.80.90	Outros	14BK

9011.90	- Partes e acessórios	
9011.90.10	Dos artigos da subposição 9011.20	0BK
9011.90.90	Outros	14BK
90.12	**Microscópios, exceto ópticos; difratógrafos.**	
9012.10	- Microscópios, exceto ópticos; difratógrafos	
9012.10.10	Microscópios eletrônicos	0BK
9012.10.90	Outros	14BK
9012.90	- Partes e acessórios	
9012.90.10	De microscópios eletrônicos	0BK
9012.90.90	Outros	14BK
90.13	**Dispositivos de cristais líquidos que não constituam artigos compreendidos mais especificamente noutras posições; lasers, exceto diodos laser; outros aparelhos e instrumentos de óptica, não especificados nem compreendidos noutras posições do presente Capítulo.**	
9013.10	- Miras telescópicas para armas; periscópios; lunetas para máquinas, aparelhos ou instrumentos do presente Capítulo ou da Seção XVI	
9013.10.10	Miras telescópicas para armas	18
9013.10.90	Outros	14BK
9013.20.00	- Lasers, exceto diodos laser	14BK
9013.80	- Outros dispositivos, aparelhos e instrumentos	
9013.80.10	Dispositivos de cristais líquidos (LCD)	0BIT
9013.80.90	Outros	18
9013.90.00	- Partes e acessórios	14BK
90.14	**Bússolas, incluindo as agulhas de marear; outros instrumentos e aparelhos de navegação.**	
9014.10.00	- Bússolas, incluindo as agulhas de marear	14BK
9014.20	- Instrumentos e aparelhos para navegação aérea ou espacial (exceto bússolas)	
9014.20.10	Altímetros	0BK
9014.20.20	Pilotos automáticos	0BK
9014.20.30	Inclinômetros	0BK
9014.20.90	Outros	0BK
9014.80	- Outros aparelhos e instrumentos	
9014.80.10	Sondas acústicas (ecobatímetros) ou de ultrassom (sonar e semelhantes)	14BK
9014.80.90	Outros	14BK
9014.90.00	- Partes e acessórios	14BK
90.15	**Instrumentos e aparelhos de geodésia, topografia, agrimensura, nivelamento, fotogrametria, hidrografia, oceanografia, hidrologia, meteorologia ou de geofísica, exceto bússolas; telêmetros.**	
9015.10.00	- Telêmetros	14BK
9015.20	- Teodolitos e taqueômetros	
9015.20.10	Com sistema de leitura por meio de prisma ou micrômetro óptico e precisão de leitura de 1 segundo	14BK
9015.20.90	Outros	14BK
9015.30.00	- Níveis	14BK
9015.40.00	- Instrumentos e aparelhos de fotogrametria	0BK

9015.80	- Outros instrumentos e aparelhos	
9015.80.10	Molinetes hidrométricos	14BK
9015.80.90	Outros	14BK
9015.90	- Partes e acessórios	
9015.90.10	De instrumentos ou aparelhos da subposição 9015.40	0BK
9015.90.90	Outros	14BK
9016.00	**Balanças sensíveis a pesos iguais ou inferiores a 5 cg, com ou sem pesos.**	
9016.00.10	Sensíveis a pesos não superiores a 0,2 mg	14BK
9016.00.90	Outras	14BK
90.17	**Instrumentos de desenho, de traçado ou de cálculo (por exemplo, máquinas de desenhar, pantógrafos, transferidores, estojos de desenho, réguas de cálculo e discos de cálculo); instrumentos de medida de distâncias de uso manual (por exemplo, metros, micrômetros, paquímetros e calibres), não especificados nem compreendidos noutras posições do presente Capítulo.**	
9017.10	- Mesas e máquinas de desenhar, mesmo automáticas	
9017.10.10	Automáticas	0BK
9017.10.90	Outras	18
9017.20.00	- Outros instrumentos de desenho, de traçado ou de cálculo	18
9017.30	- Micrômetros, paquímetros, calibres e semelhantes	
9017.30.10	Micrômetros	10
9017.30.20	Paquímetros	18
9017.30.90	Outros	18
9017.80	- Outros instrumentos	
9017.80.10	Metros	18
9017.80.90	Outros	18
9017.90	- Partes e acessórios	
9017.90.10	De mesas ou máquinas de desenhar, automáticas	0BK
9017.90.90	Outros	16
90.18	**Instrumentos e aparelhos para medicina, cirurgia, odontologia e veterinária, incluindo os aparelhos para cintilografia e outros aparelhos eletromédicos, bem como os aparelhos para testes visuais.**	
9018.1	- Aparelhos de eletrodiagnóstico (incluindo os aparelhos de exploração funcional e os de verificação de parâmetros fisiológicos):	
9018.11.00	-- Eletrocardiógrafos	14BK
9018.12	-- Aparelhos de diagnóstico por varredura ultrassônica (*scanners*)	
9018.12.10	Ecógrafos com análise espectral Doppler	0BK
9018.12.90	Outros	14BK
9018.13.00	-- Aparelhos de diagnóstico por visualização de ressonância magnética	0BK
9018.14	-- Aparelhos de cintilografia	
9018.14.10	*Scanner* de tomografia por emissão de posítrons (PET - *Positron Emission Tomography*)	0BK
9018.14.20	Câmaras gama	0BK
9018.14.90	Outros	14BK
9018.19	-- Outros	
9018.19.10	Endoscópios	0BK
9018.19.20	Audiômetros	14BK

9018.19.80	Outros	14BK
9018.19.90	Partes	14BK
9018.20	- Aparelhos de raios ultravioleta ou infravermelhos	
9018.20.10	Para cirurgia, que operem por laser	0BK
9018.20.20	Outros, para tratamento bucal, que operem por laser	0BK
9018.20.90	Outros	14BK
9018.3	- Seringas, agulhas, cateteres, cânulas e instrumentos semelhantes:	
9018.31	-- Seringas, mesmo com agulhas	
9018.31.1	De plástico	
9018.31.11	De capacidade inferior ou igual a 2 cm^3	16
9018.31.19	Outras	16
9018.31.90	Outras	16
9018.32	-- Agulhas tubulares de metal e agulhas para suturas	
9018.32.1	Tubulares de metal	
9018.32.11	Gengivais	16
9018.32.12	De aço cromo-níquel, bisel trifacetado e diâmetro exterior superior ou igual a 1,6 mm, do tipo das utilizadas com bolsas de sangue	2
9018.32.19	Outras	16
9018.32.20	Para suturas	2
9018.39	-- Outros	
9018.39.10	Agulhas	16
9018.39.2	Sondas, cateteres e cânulas	
9018.39.21	De borracha	16#
9018.39.22	Cateteres de poli(cloreto de vinila), para embolectomia arterial	2
9018.39.23	Cateteres de poli(cloreto de vinila), para termodiluição	2
9018.39.24	Cateteres intravenosos periféricos, de poliuretano ou de copolímero de etileno-tetrafluoretileno (ETFE)	16
9018.39.29	Outros	16#
9018.39.30	Lancetas para vacinação e cautérios	16
9018.39.9	Outros	
9018.39.91	Artigo para fístula arteriovenosa, composto de agulha, base de fixação tipo borboleta, tubo plástico com conector e obturador	16
9018.39.99	Outros	16
9018.4	- Outros instrumentos e aparelhos para odontologia:	
9018.41.00	-- Aparelhos dentários de brocar, mesmo combinados numa base comum com outros equipamentos dentários	14BK
9018.49	-- Outros	
9018.49.1	Brocas	
9018.49.11	De carboneto de tungstênio (volfrâmio)	16
9018.49.12	De aço-vanádio	16
9018.49.19	Outras	16
9018.49.20	Limas	16
9018.49.40	Para tratamento bucal, que operem por projeção cinética de partículas	0BK
9018.49.9	Outros	
9018.49.91	Para desenho e construção de peças cerâmicas para restaurações dentárias, computadorizados	0BK
9018.49.99	Outros	16
9018.50	- Outros instrumentos e aparelhos para oftalmologia	
9018.50.10	Microscópios binoculares, dos tipos utilizados em cirurgia oftalmológica	0BK
9018.50.90	Outros	14BK

9018.90	- Outros instrumentos e aparelhos	
9018.90.10	Para transfusão de sangue ou infusão intravenosa	14BK
9018.90.2	Bisturis	
9018.90.21	Elétricos	16
9018.90.29	Outros	16
9018.90.3	Litótomos e litotritores	
9018.90.31	Litotritores por onda de choque	0BK
9018.90.39	Outros	14BK
9018.90.40	Rins artificiais	8BK
9018.90.50	Aparelhos de diatermia	14BK
9018.90.9	Outros	
9018.90.91	Incubadoras para bebês	14BK
9018.90.92	Aparelhos para medida da pressão arterial	16
9018.90.93	Aparelhos para terapia intra-uretral por micro-ondas (TUMT), próprios para o tratamento de afecções prostáticas, computadorizados	0BK
9018.90.94	Endoscópios	0BK
9018.90.95	Grampos e clipes, seus aplicadores e extratores	0
9018.90.96	Desfibriladores externos que operem unicamente em modo automático (AED - *Automatic External Defibrillator*)	0
9018.90.99	Outros	16#
90.19	**Aparelhos de mecanoterapia; aparelhos de massagem; aparelhos de psicotécnica; aparelhos de ozonoterapia, de oxigenoterapia, de aerossolterapia, aparelhos respiratórios de reanimação e outros aparelhos de terapia respiratória.**	
9019.10.00	- Aparelhos de mecanoterapia; aparelhos de massagem; aparelhos de psicotécnica	14BK
9019.20	- Aparelhos de ozonoterapia, de oxigenoterapia, de aerossolterapia, aparelhos respiratórios de reanimação e outros aparelhos de terapia respiratória	
9019.20.10	De oxigenoterapia	14BK
9019.20.20	De aerossolterapia	14BK
9019.20.30	Respiratórios de reanimação	14BK
9019.20.40	Respiradores automáticos (pulmões de aço)	14BK
9019.20.90	Outros	14BK
9020.00	**Outros aparelhos respiratórios e máscaras contra gases, exceto as máscaras de proteção desprovidas de mecanismo e de elemento filtrante amovível.**	
9020.00.10	Máscaras contra gases	16
9020.00.90	Outros	16
90.21	**Artigos e aparelhos ortopédicos, incluindo as cintas e fundas médico-cirúrgicas e as muletas; talas, goteiras e outros artigos e aparelhos para fraturas; artigos e aparelhos de prótese; aparelhos para facilitar a audição dos surdos e outros aparelhos para compensar deficiências ou enfermidades, que se destinam a ser transportados a mão ou sobre as pessoas ou a ser implantados no organismo.**	
9021.10	- Artigos e aparelhos ortopédicos ou para fraturas	
9021.10.10	Artigos e aparelhos ortopédicos	14
9021.10.20	Artigos e aparelhos para fraturas	14#
9021.10.9	Partes e acessórios	

9021.10.91	De artigos e aparelhos de ortopedia, articulados	0
9021.10.99	Outros	14
9021.2	- Artigos e aparelhos de prótese dentária:	
9021.21	-- Dentes artificiais	
9021.21.10	De acrílico	16
9021.21.90	Outros	16
9021.29.00	-- Outros	16
9021.3	- Outros artigos e aparelhos de prótese:	
9021.31	-- Próteses articulares	
9021.31.10	Femurais	14#
9021.31.20	Mioelétricas	0
9021.31.90	Outras	14#
9021.39	-- Outros	
9021.39.1	Válvulas cardíacas	
9021.39.11	Mecânicas	0
9021.39.19	Outras	14
9021.39.20	Lentes intraoculares	0
9021.39.30	Próteses de artérias vasculares revestidas	0
9021.39.40	Próteses mamárias não implantáveis	0
9021.39.80	Outros	14
9021.39.9	Partes e acessórios	
9021.39.91	Partes de próteses modulares que substituem membros superiores ou inferiores	0
9021.39.99	Outros	14
9021.40.00	- Aparelhos para facilitar a audição dos surdos, exceto as partes e acessórios	0
9021.50.00	- Marca-passos cardíacos, exceto as partes e acessórios	14#
9021.90	- Outros	
9021.90.1	Aparelhos que se implantam no organismo para compensar um defeito ou uma incapacidade	
9021.90.11	Cardiodesfibriladores automáticos	0
9021.90.19	Outros	0
9021.90.8	Outros	
9021.90.81	Implantes expansíveis (*stents*), mesmo montados sobre cateter do tipo balão	0
9021.90.82	Oclusores interauriculares constituídos por uma malha de fios de níquel e titânio preenchida com tecido de poliéster, mesmo apresentados com seu respectivo cateter	0
9021.90.89	Outros	14
9021.90.9	Partes e acessórios	
9021.90.91	De marca-passos cardíacos	0
9021.90.92	De aparelhos para facilitar a audição dos surdos	0
9021.90.99	Outros	14
90.22	**Aparelhos de raios X e aparelhos que utilizem radiações alfa, beta ou gama, mesmo para usos médicos, cirúrgicos, odontológicos ou veterinários, incluindo os aparelhos de radiofotografia ou de radioterapia, os tubos de raios X e outros dispositivos geradores de raios X, os geradores de tensão, as mesas de comando, as telas de visualização, as mesas, poltronas e suportes semelhantes para exame ou tratamento.**	

9022.1	- Aparelhos de raios X, mesmo para usos médicos, cirúrgicos, odontológicos ou veterinários, incluindo os aparelhos de radiofotografia ou de radioterapia:	
9022.12.00	-- Aparelhos de tomografia computadorizada	0BK
9022.13	-- Outros, para odontologia	
9022.13.1	De diagnóstico	
9022.13.11	De tomadas maxilares panorâmicas	0BK
9022.13.19	Outros	14BK
9022.13.90	Outros	0BK
9022.14	-- Outros, para usos médicos, cirúrgicos ou veterinários	
9022.14.1	De diagnóstico	
9022.14.11	Para mamografia	14BK
9022.14.12	Para angiografia	0BK
9022.14.13	Para densitometria óssea, computadorizados	0BK
9022.14.19	Outros	14BK
9022.14.90	Outros	0BK
9022.19	-- Para outros usos	
9022.19.10	Espectrômetros ou espectrógrafos de raios X	0BK
9022.19.9	Outros	
9022.19.91	Dos tipos utilizados para inspeção de bagagens, com túnel de altura inferior ou igual a 0,4 m, largura inferior ou igual a 0,6 m e comprimento inferior ou igual a 1,2 m	14BK
9022.19.99	Outros	0BK
9022.2	- Aparelhos que utilizem radiações alfa, beta ou gama, mesmo para usos médicos, cirúrgicos, odontológicos ou veterinários, incluindo os aparelhos de radiofotografia ou de radioterapia:	
9022.21	-- Para usos médicos, cirúrgicos, odontológicos ou veterinários	
9022.21.10	Aparelhos de radiocobalto (bombas de cobalto)	14BK
9022.21.20	Outros, para gamaterapia	0BK
9022.21.90	Outros	0BK
9022.29	-- Para outros usos	
9022.29.10	Para detecção do nível de enchimento ou tampas faltantes, em latas de bebidas, por meio de raios gama	0BK
9022.29.90	Outros	14BK
9022.30.00	- Tubos de raios X	0BK
9022.90	- Outros, incluindo as partes e acessórios	
9022.90.1	Aparelhos	
9022.90.11	Geradores de tensão	14BK
9022.90.12	Telas radiológicas	14BK
9022.90.19	Outros	14BK
9022.90.80	Outros	14BK
9022.90.90	Partes e acessórios de aparelhos de raios X	14BK
9023.00.00	**Instrumentos, aparelhos e modelos, concebidos para demonstração (por exemplo, no ensino e nas exposições), não suscetíveis de outros usos.**	16
90.24	**Máquinas e aparelhos para ensaios de dureza, tração, compressão, elasticidade ou de outras propriedades mecânicas de materiais (por exemplo, metais, madeira, têxteis, papel, plásticos).**	
9024.10	- Máquinas e aparelhos para ensaios de metais	
9024.10.10	Para ensaios de tração ou compressão	14BK

9024.10.20	Para ensaios de dureza	14BK
9024.10.90	Outros	14BK
9024.80	- Outras máquinas e aparelhos	
9024.80.1	Máquinas e aparelhos para ensaios de têxteis	
9024.80.11	Automáticos, para fios	0BK
9024.80.19	Outros	14BK
9024.80.2	Máquinas e aparelhos para ensaios de papel, cartão, linóleo e plástico ou borracha flexíveis	
9024.80.21	Máquinas para ensaios de pneumáticos	0BK
9024.80.29	Outros	14BK
9024.80.90	Outros	14BK
9024.90.00	- Partes e acessórios	14BK
90.25	**Densímetros, areômetros, pesa-líquidos e instrumentos flutuantes semelhantes, termômetros, pirômetros, barômetros, higrômetros e psicrômetros, registradores ou não, mesmo combinados entre si.**	
9025.1	- Termômetros e pirômetros, não combinados com outros instrumentos:	
9025.11	-- De líquido, de leitura direta	
9025.11.10	Termômetros clínicos	18
9025.11.90	Outros	18
9025.19	-- Outros	
9025.19.10	Pirômetros ópticos	0
9025.19.90	Outros	18
9025.80.00	- Outros instrumentos	18
9025.90	- Partes e acessórios	
9025.90.10	De termômetros	16
9025.90.90	Outros	16
90.26	**Instrumentos e aparelhos para medida ou controle da vazão, do nível, da pressão ou de outras características variáveis dos líquidos ou gases (por exemplo, medidores de vazão, indicadores de nível, manômetros, contadores de calor), exceto os instrumentos e aparelhos das posições 90.14, 90.15, 90.28 ou 90.32.**	
9026.10	- Para medida ou controle da vazão ou do nível dos líquidos	
9026.10.1	Para medida ou controle de vazão	
9026.10.11	Medidores-transmissores eletrônicos, que funcionem pelo princípio de indução eletromagnética	14BIT
9026.10.19	Outros	18
9026.10.2	Para medida ou controle do nível	
9026.10.21	De metais, mediante correntes parasitas	2
9026.10.29	Outros	18
9026.20	- Para medida ou controle da pressão	
9026.20.10	Manômetros	18
9026.20.90	Outros	18
9026.80.00	- Outros instrumentos e aparelhos	18
9026.90	- Partes e acessórios	
9026.90.10	De instrumentos e aparelhos para medida ou controle do nível	16
9026.90.20	De manômetros	16
9026.90.90	Outros	16

90.27	**Instrumentos e aparelhos para análises físicas ou químicas (por exemplo, polarímetros, refratômetros, espectrômetros, analisadores de gases ou de fumaça); instrumentos e aparelhos para ensaios de viscosidade, porosidade, dilatação, tensão superficial ou semelhantes ou para medidas calorimétricas, acústicas ou fotométricas (incluindo os indicadores de tempo de exposição); micrótomos.**	
9027.10.00	- Analisadores de gases ou de fumaça	14BK
9027.20	- Cromatógrafos e aparelhos de eletroforese	
9027.20.1	Cromatógrafos	
9027.20.11	De fase gasosa	0BK
9027.20.12	De fase líquida	0BK
9027.20.19	Outros	0BK
9027.20.2	Aparelhos de eletroforese	
9027.20.21	Sequenciadores automáticos de ADN mediante eletroforese capilar	0BK
9027.20.29	Outros	14BK
9027.30	- Espectrômetros, espectrofotômetros e espectrógrafos que utilizem radiações ópticas (UV, visíveis, IV)	
9027.30.1	Espectrômetros e espectrógrafos	
9027.30.11	De emissão atômica	0BK
9027.30.19	Outros	14BK
9027.30.20	Espectrofotômetros	14BK
9027.50	- Outros aparelhos e instrumentos que utilizem radiações ópticas (UV, visíveis, IV)	
9027.50.10	Colorímetros	14BK
9027.50.20	Fotômetros	14BK
9027.50.30	Refratômetros	14BK
9027.50.40	Sacarímetros	14BK
9027.50.50	Citômetro de fluxo	0BK
9027.50.90	Outros	14BK
9027.80	- Outros instrumentos e aparelhos	
9027.80.1	Calorímetros, viscosímetros, densitômetros e aparelhos medidores de pH	
9027.80.11	Calorímetros	0BK
9027.80.12	Viscosímetros	14BK
9027.80.13	Densitômetros	14BK
9027.80.14	Aparelhos medidores de pH	14BK
9027.80.20	Espectrômetros de massa	14BK
9027.80.30	Polarógrafos	0BK
9027.80.9	Outros	
9027.80.91	Exposímetros	0BK
9027.80.99	Outros	14BK
9027.90	- Micrótomos; partes e acessórios	
9027.90.10	Micrótomos	0BK
9027.90.9	Partes e acessórios	
9027.90.91	De espectrômetros e espectrógrafos, de emissão atômica	0BK
9027.90.93	De polarógrafos	0BK
9027.90.99	Outros	14BK
90.28	**Contadores de gases, de líquidos ou de eletricidade, incluindo os aparelhos para sua aferição.**	
9028.10	- Contadores de gases	
9028.10.1	De gás natural comprimido, eletrônicos	

9028.10.11	Dos tipos utilizados em postos (estações) de serviço ou garagens	14BK
9028.10.19	Outros	14BK
9028.10.90	Outros	18
9028.20	- Contadores de líquidos	
9028.20.10	De peso inferior ou igual a 50 kg	18
9028.20.20	De peso superior a 50 kg	18
9028.30	- Contadores de eletricidade	
9028.30.1	Monofásicos, para corrente alternada	
9028.30.11	Digitais	14BIT
9028.30.19	Outros	18
9028.30.2	Bifásicos	
9028.30.21	Digitais	14BIT
9028.30.29	Outros	18
9028.30.3	Trifásicos	
9028.30.31	Digitais	14BIT
9028.30.39	Outros	18
9028.30.90	Outros	18
9028.90	- Partes e acessórios	
9028.90.10	De contadores de eletricidade	16
9028.90.90	Outros	16
90.29	**Outros contadores (por exemplo, contadores de voltas, contadores de produção, taxímetros, totalizadores de caminho percorrido, podômetros); indicadores de velocidade e tacômetros, exceto os das posições 90.14 ou 90.15; estroboscópios.**	
9029.10	- Contadores de voltas, contadores de produção, taxímetros, totalizadores de caminho percorrido, podômetros e contadores semelhantes	
9029.10.10	Contadores de voltas, contadores de produção ou de horas de trabalho	14BK
9029.10.90	Outros	18
9029.20	- Indicadores de velocidade e tacômetros; estroboscópios	
9029.20.10	Indicadores de velocidade e tacômetros	18
9029.20.20	Estroboscópios	18
9029.90	- Partes e acessórios	
9029.90.10	De indicadores de velocidade e tacômetros	16
9029.90.90	Outros	16
90.30	**Osciloscópios, analisadores de espectro e outros instrumentos e aparelhos para medida ou controle de grandezas elétricas; instrumentos e aparelhos para medida ou detecção de radiações alfa, beta, gama, X, cósmicas ou outras radiações ionizantes.**	
9030.10	- Instrumentos e aparelhos para medida ou detecção de radiações ionizantes	
9030.10.10	Medidores de radioatividade	14BK
9030.10.90	Outros	14BK
9030.20	- Osciloscópios e oscilógrafos	
9030.20.10	Osciloscópios digitais	2§BIT
9030.20.2	Osciloscópios analógicos	
9030.20.21	De frequência superior ou igual a 60 MHz	2§BIT
9030.20.22	Vetorscópios	2§BIT
9030.20.29	Outros	12BIT
9030.20.30	Oscilógrafos	0BK

9030.3	- Outros aparelhos e instrumentos para medida ou controle da tensão, intensidade, resistência ou da potência:	
9030.31.00	-- Multímetros, sem dispositivo registrador	14BK
9030.32.00	-- Multímetros, com dispositivo registrador	14BK
9030.33	-- Outros, sem dispositivo registrador	
9030.33.1	Voltímetros	
9030.33.11	Digitais	12BIT
9030.33.19	Outros	12BIT
9030.33.2	Amperímetros	
9030.33.21	Do tipo dos utilizados em veículos automóveis	18
9030.33.29	Outros	14BK
9030.33.90	Outros	14BK
9030.39	-- Outros, com dispositivo registrador	
9030.39.10	De teste de continuidade em circuitos impressos	12BIT
9030.39.90	Outros	14BK
9030.40	- Outros instrumentos e aparelhos, especialmente concebidos para telecomunicações (por exemplo, diafonômetros, medidores de ganho, distorciômetros, psofômetros)	
9030.40.10	Analisadores de protocolo	12BIT
9030.40.20	Analisadores de nível seletivo	12BIT
9030.40.30	Analisadores digitais de transmissão	12BIT
9030.40.90	Outros	12BIT
9030.8	- Outros instrumentos e aparelhos:	
9030.82	-- Para medida ou controle de plaquetas (*wafers*) ou de dispositivos semicondutores	
9030.82.10	De testes de circuitos integrados	12BIT
9030.82.90	Outros	12BIT
9030.84	-- Outros, com dispositivo registrador	
9030.84.10	De teste automático de circuito impresso montado (ATE)	2§BIT
9030.84.20	De medidas de parâmetros característicos de sinais de televisão ou de vídeo	2§BIT
9030.84.90	Outros	14BK
9030.89	-- Outros	
9030.89.10	Analisadores lógicos de circuitos digitais	2§BIT
9030.89.20	Analisadores de espectro de frequência	2§BIT
9030.89.30	Frequencímetros	12BIT
9030.89.40	Fasímetros	12BIT
9030.89.90	Outros	12BIT
9030.90	- Partes e acessórios	
9030.90.10	De instrumentos e aparelhos da subposição 9030.10	14BK
9030.90.90	Outros	8BIT
90.31	**Instrumentos, aparelhos e máquinas de medida ou controle, não especificados nem compreendidos noutras posições do presente Capítulo; projetores de perfis.**	
9031.10.00	- Máquinas de balancear (equilibrar) peças mecânicas	14BK
9031.20	- Bancos de ensaio	
9031.20.10	Para motores	14BK
9031.20.90	Outros	14BK
9031.4	- Outros instrumentos e aparelhos ópticos:	

9031.41.00	-- Para controle de plaquetas (*wafers*) ou de dispositivos semicondutores ou para controle de máscaras ou retículos utilizados na fabricação de dispositivos semicondutores	14BK
9031.49	-- Outros	
9031.49.10	Para medida de parâmetros dimensionais de fibras de celulose, por meio de raios laser	0BK
9031.49.20	Para medida da espessura de pneumáticos de veículos automóveis, por meio de raios laser	0BK
9031.49.90	Outros	14BK
9031.80	- Outros instrumentos, aparelhos e máquinas	
9031.80.1	Dinamômetros e rugosímetros	
9031.80.11	Dinamômetros	14BK
9031.80.12	Rugosímetros	14BK
9031.80.20	Máquinas para medição tridimensional	14BK
9031.80.30	Metros padrões	10BK
9031.80.40	Aparelhos digitais, de uso em veículos automóveis, para medida e indicação de múltiplas grandezas tais como: velocidade média, consumos instantâneo e médio e autonomia (computador de bordo)	16BIT
9031.80.50	Aparelhos para análise de têxteis, computadorizados	0BK
9031.80.60	Células de carga	14BK
9031.80.9	Outros	
9031.80.91	Para controle dimensional de pneumáticos, em condições de carga	0BK
9031.80.99	Outros	14BK
9031.90	- Partes e acessórios	
9031.90.10	De bancos de ensaio	14BK
9031.90.90	Outros	14BK
90.32	**Instrumentos e aparelhos para regulação ou controle, automáticos.**	
9032.10	- Termostatos	
9032.10.10	De expansão de fluidos	18
9032.10.90	Outros	18
9032.20.00	- Manostatos (pressostatos)	18
9032.8	- Outros instrumentos e aparelhos:	
9032.81.00	-- Hidráulicos ou pneumáticos	18
9032.89	-- Outros	
9032.89.1	Reguladores de voltagem	
9032.89.11	Eletrônicos	12BIT
9032.89.19	Outros	18
9032.89.2	Controladores eletrônicos do tipo dos utilizados em veículos automóveis	
9032.89.21	De sistemas antibloqueantes de freio (ABS)	16BIT
9032.89.22	De sistemas de suspensão	16BIT
9032.89.23	De sistemas de transmissão	16BIT
9032.89.24	De sistemas de ignição	16BIT
9032.89.25	De sistemas de injeção	16BIT
9032.89.29	Outros	16BIT
9032.89.30	Equipamentos digitais para controle de veículos ferroviários	14BIT
9032.89.8	Outros, para regulação ou controle de grandezas não elétricas	
9032.89.81	De pressão	14BIT
9032.89.82	De temperatura	14BIT
9032.89.83	De umidade	14BIT

9032.89.84	De velocidade de motores elétricos por variação de frequência	14BK
9032.89.89	Outros	14BIT
9032.89.90	Outros	18
9032.90	- Partes e acessórios	
9032.90.10	Circuitos impressos com componentes elétricos ou eletrônicos, montados	12BIT
9032.90.9	Outros	
9032.90.91	De termostatos	16
9032.90.99	Outros	8BIT
9033.00.00	**Partes e acessórios não especificados nem compreendidos noutras posições do presente Capítulo, para máquinas, aparelhos, instrumentos ou artigos do Capítulo 90.**	16